SEQUENTIAL DESIGN OF EXPERIMENTS FOR ANOMALY DETECTION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Chao Wang

May 2018

SEQUENTIAL DESIGN OF EXPERIMENTS FOR ANOMALY DETECTION

Chao Wang, Ph.D.

Cornell University 2018

This dissertation focuses on the sequential design of experiments for anomaly detection. Specifically, the problem of detecting a few anomalous processes among a large number of processes is considered. The rare events may represent opportunities with exceptional returns or anomalies associated with high costs or potential catastrophic consequences. Examples include financial trading opportunities and transmission opportunities in dynamic spectrum access, endogenous extreme events or exogenous attacks in communication and computer networks, etc.

For all these applications, the problem of searching for the rare has the following defining features: (i) the massive search space; (ii) the need for high detection accuracy, especially in terms of missing a rare event; (iii) the time sensitivity of the problem, either due to the transient nature of opportunities or the urgency for taking recourse measures in the face of anomalies. The goal is thus to detect the rare events as quickly and as reliably as possible when the total number of hypotheses is large and the observations are probabilistic thus inherently ambiguous. The performance measure of interest is sample complexity (the total number of observations which represents the detection delay) with respect to the size of the search space and the required detection accuracy. The key to a sublinear scaling in the problem size is to exploit the hierarchical structure of the search space inherent to many applications.

In the first part of the dissertation, we develop an algorithm for the anomaly detection of which the sample complexity is in optimal scaling with the size of the search space. We consider the case where the observations from all the processes are noiseless.

The anomaly detection problem falls into the general class of the group testing problem. We consider the quantitative group testing problem where the objective is to identify defective items in a given population based on results of tests performed on subsets of the population. Under the quantitative group testing model, the result of each test reveals the number of defective items in the tested group. We establish the optimal nested test plan in closed form which achieves the minimum number of tests by nested test plans. This optimal nested test plan is also order-optimal among all test plans as the population size approaches infinity. Using heavy-hitter detection as a case study, we show via simulation examples orders of magnitude improvement of the group testing approach over two prevailing sampling-based approaches in detection accuracy and counter consumption.

In the second part of the dissertation, we develop an algorithm for the anomaly detection problem of which the sample complexity achieves optimal scaling with the size of the search space as well as the accuracy requirements. We consider the case where the observations from the processes are noisy and the noisy observations are specific by general distributions. Aggregated observations can be taken from a chosen subset of processes, where the chosen subset conforms to a binary tree structure. The random observations are drawn from a general distribution that may depend on the size of the chosen subset and the number of anomalous processes in the subset. We propose a sequential search strategy by devising an information-directed random walk (IRW) on the tree-structured observation hierarchy. Subject to a reliable constraint, the proposed policy is shown to be asymptotically optimal in terms of detection accuracy. Furthermore, it achieves the optimal logarithmic-order sample complexity in the in terms of the size of the search space provided that the Kullback-Liebler divergence between aggregated observations in the presence and the absence of anomalous processes are bounded away from zero at all levels of the tree structure as the size of the search space approaches infinity. Sufficient conditions on the decaying rate of the aggregated observations to

pure noise under which a sublinear scaling in the size of the search space is preserved are also identified for the Bernoulli case.

The algorithms proposed in both of the two parts are adaptive test plans which are deterministic with search actions explicitly specified at each given time. They involve little online computation beyond calculating the sample mean or the sum log-likelihood ratio. The inherent tree structure of the also leads to low memory requirement. They are thus particularly attractive for online applications.

BIOGRAPHICAL SKETCH

Chao Wang received the B.S. degree in physics from University of Science and Technology of China (USTC), Hefei, China, and the M.S. degree in electrical and computer engineering from University of California, Davis, CA, USA, in 2012 and 2014, respectively. He will receive the Ph.D. degree from Cornell University, Ithaca, NY, USA, in 2018. His research interests include the active learning, decision theory, and signal detection and estimation.

To my family and friends.

ACKNOWLEDGEMENTS

First and foremost, I wish to express my deepest gratitude to my advisor, Professor Qing Zhao, for her continuous support, encouragement, guidance and countless help for my Ph.D. study throughout all these years. I have been deeply stimulated and influenced by Professor Zhaos integrity and enthusiasm to scientific and engineering research. I have benefited and learned a lot from research discussions with her and class lectures given by her. It is my great honor to be her student.

I would like to acknowledge Professor A. Kevin Tang and Professor Lang Tong for serving on my Ph.D. committee in Cornell University. I would also like to extend my gratitude to Professor Chen-Nee Chuah and Professor Bernard C. Levy in University of California, Davis for serving my Ph.D. committee before I transferred to Cornell. The discussions and suggestions from all of them are very informative and help me to deepen my understanding of the research problems.

Special thanks go to Dr. Kobi Cohen for the discussions and suggestions on the research projects. Those insightful comments and discussions are extremely helpful in tackling these hard problems.

I would also like to express my gratitude to my former and current colleagues and labmates: Pouya Tehrani, Jianhang Gao, Yiyxuan Zhai, Yuan Zhou, Sattar Vakili, Xiao Xu, Boshuang Huang, Professor Guanghua Song and Professor Hui Feng. It has been a great pleasure working with all of them.

I also want to thank my parents. They have always been my strongest backing during my whole life. Their wholehearted support makes it possible for me to focus on my Ph.D. research.

There are no words to convey how much I appreciate and love my girlfriend, soulmate, Ming Chen. Were it not for her unreserved dedication and unconditional love, I could not have finished my Ph.D. study and this dissertation. She has always been

understanding and supportive during my good and bad times.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgements . . . v
Table of Contents . . . vii
List of Tables . . . x
List of Figures . . . xi

1 Introduction **1**
1.1 Searching for the Rare . . . 1
1.2 Optimal Scaling with the Size of the Search Space . . . 3
1.2.1 Group Testing Problem . . . 3
1.2.2 Main Results . . . 4
1.3 Optimal Scaling with the Size of the Search Space and Accuracy in Noisy Scenarios . . . 6
1.3.1 Active Hypothesis Testing . . . 7
1.3.2 Main Results . . . 8
1.4 Organization . . . 9

2 Achieving Optimal Scaling with the Size of the Search Space **10**
2.1 Background and Related Work . . . 10
2.1.1 Classic Group Testing . . . 10
2.1.2 Quantitative Group Testing . . . 11
2.1.3 Related Work . . . 12
2.2 Problem Formulation . . . 14
2.3 The Optimal Nested Test Plan . . . 17
2.3.1 $N(n,d)$ and Its Geometric Block-Constant Structure . . . 18
2.3.2 The Optimal Nested Test Plan . . . 20
2.3.3 The Optimal Nested Test Plan for CGT with Unknown d . . . 22
2.3.4 Order Optimality and the Approximation Ratio of the Optimal Nested Test Plan . . . 23
2.4 Properties of $N(n,d)$ and Proof of Theorem 1 . . . 25
2.4.1 Properties of $N(n,d)$ . . . 25
2.4.2 Proof of Theorem 1 . . . 26
2.5 Comparison between Quantitative and Boolean Group Testing . . . 28
2.5.1 Comparison for Cases with Known d . . . 28
2.5.2 Comparison for Cases with Unknown d . . . 30
2.6 Application to Heavy Hitter Detection . . . 30
2.6.1 Quantitative Group Testing for Heavy Hitter Detection . . . 32
2.6.2 Comparisons with Prevailing Heavy Hitter Detectors . . . 34

3 Achieving the Optimal Scaling with the Size of the Search Space and Accuracy in Noisy Scenarios **41**
3.1 Background and Related Work . . . 41
3.1.1 Active Hypothesis Testing . . . 41
3.1.2 Related Work . . . 42
3.2 Problem Formulation . . . 45
3.3 Information-Directed Random Walk . . . 47
3.3.1 The Global Random Walk Module . . . 48
3.3.2 The Local Test Module . . . 49
3.4 Performance Analysis of IRW Policy . . . 51
3.4.1 Main Structure of the Analysis . . . 51
3.4.2 Informative Observations at All Levels . . . 53
3.4.3 Aggregated Observations Decaying to Pure Noise . . . 54
3.5 Multiple Targets and General Tree Structures . . . 56
3.5.1 Formulation of Multiple Target Detection . . . 56
3.5.2 IRW Policy for Known L . . . 57
3.5.3 IRW Policy for Unknown L . . . 60
3.5.4 General Tree Structures . . . 62
3.6 Discussions . . . 64
3.6.1 Channel Coding with Feedback . . . 64
3.6.2 Noisy Group Testing and Compressed Sensing . . . 68
3.6.3 Adaptive Sampling with Noisy Response . . . 69
3.7 Simulation Examples . . . 71

4 Conclusion **76**

A Proof for Lemmas and Theorems in Chapter 2 **77**
A.1 Proof of Lemma 1 . . . 77
A.2 Proof of Properties of $N(n,d)$ . . . 78
A.2.1 Proof of [P1] . . . 78
A.2.2 Proof of [P2] . . . 81
A.2.3 Proof of [P3] . . . 87

B Proof for Theorems in Chapter 3 **88**
B.1 Two Sequential Versions of the Local Tests . . . 88
B.1.1 Passive Sequential Local Test . . . 88
B.1.2 Active Sequential Local Test . . . 89
B.1.3 Sequential Local Tests for Multiple Targets Detection . . . 92
B.2 Proof of Theorem 3 . . . 93
B.3 Proof of Theorem 4 and Theorem 5 . . . 98
B.4 Proof of Theorem 6 . . . 100
B.4.1 Detection delay without detection errors on the tree . . . 101
B.4.2 Detection delay with detection errors on the tree . . . 103

Bibliography **109**

LIST OF TABLES

2.1 The Frame-Segment Structure of $N(n,d)$ 19
2.2 The Frame-Segment Structure of $M(n,d)$ 20
2.3 A comparative summary of boolean and quantitative CGT results. . . . 28

LIST OF FIGURES

2.1 Comparison of the optimal nested test plan to the generalized binary splitting (GBS) test plan with known d ($n = 500$, 1000 Monte Carlo runs). 31
2.2 Comparison of the optimal nested test plan to the binary splitting test plan with unknown d ($n = 500$, 1000 Monte Carlo runs). 31
2.3 Detection accuracy of the optimal nested test plan with MLE for Poisson distributed flows ($n = 1000$, $T = 2$, $\lambda_y = 1$). 35
2.4 Detection accuracy of the optimal nested test plan with SME for log-normal distributed flows ($n = 1000$, $T = 5$, $\lambda_y = 1$, $\sigma_x^2 = \sigma_y^2 = 10$). . . . 36
2.5 Performance comparison: detection accuracy versus detection delay ($n = 100$ Poisson flows, $n_x = 3$, $\lambda_x = 20$, $\lambda_y = 1$, $c = 3$). 38
2.6 Performance comparison: detection accuracy versus counter budget ($n = 1000$ Poisson flows, $n_x = 200$, $\lambda_x = 36$, $\lambda_y = 1$, $\tau = 568$). 39

3.1 A binary tree observation model with a single target. 47
3.2 A biased random walk on the tree. 50
3.3 A biased random walk on the tree with sojourn times at the leaves when there is a single target. 53
3.4 A binary tree observation model with multiple targets. 56
3.5 A biased random walk on the tree with sojourn times at the leaves when there are multiple targets. 61
3.6 A general tree with bounded degree. 63
3.7 A channel coding with noiseless feedback. 65
3.8 Root mean squared error of IRW for point transmission through a BSC. 68
3.9 Performance comparison of three test plans ($L = 1$, $\lambda_g = 10, \lambda_f = 0.01, K_l = 3, c = 10^{-13}$, $n = 8, 16, \ldots, 1024$.). 73
3.10 Performance comparison of IRW with three different local tests. ($K_l = 7$, $\gamma_1 = 1.0986$, $\gamma_0 = -1.0986$, $\nu_1 = 0.9445$, $\nu_0 = -0.9445$, and 1000 Monte Carlo runs.) 74
3.11 Performance comparison of three test plans ($L = 5$, $\lambda_g = 10, \lambda_f = 0.001, c = 5 \times 10^{-5}$, $n = 8, 16, \ldots, 1024$.). 75

A.1 Illustration of Lemma 2 with $s = 3$. 83

B.1 A biased random walk on the tree with detection errors. 104
B.2 A biased random walk on the tree with detection errors: nested affected trees. 108

CHAPTER 1

INTRODUCTION

1.1 Searching for the Rare

The problem of searching for a few rare events of interest among a massive number of possibilities is ubiquitous. The rare events may represent opportunities with exceptional returns or anomalies associated with high costs or potential catastrophic consequences. Examples include financial trading opportunities and transmission opportunities in dynamic spectrum access, endogenous extreme events or exogenous attacks in communication and computer networks, structural anomalies on bridges or buildings, and high-risk contingencies in power systems that may lead to cascading failures.

Regardless of the application domain, the problem of searching for the rare has the following defining features: (i) the massive search space; (ii) the need for high detection accuracy, especially in terms of missing a rare event; (iii) the time sensitivity of the problem, either due to the transient nature of opportunities or the urgency for taking recourse measures in the face of anomalies. The goal is thus to detect the rare events as quickly and as reliably as possible when the total number of hypotheses is large and the observations are probabilistic thus inherently ambiguous. The performance measure of interest is sample complexity (the total number of observations which represents the detection delay) with respect to the size of the search space and the required detection accuracy.

A question of particular interest is whether a sublinear scaling of the sample complexity with respect to the search space is feasible while achieving the optimal scaling with respect to detection accuracy. In other words, whether accurate detection can be

achieved by examining only a diminishing fraction of the search space as the search space grows.

The key to a sublinear scaling in the problem size is to exploit the hierarchical structure of the search space inherent to many applications. For example, financial transactions can be aggregated at different temporal and geographic scales. In computer vision applications such as bridge inspection by UAVs with limited battery capacity, sequentially determining areas to zoom in or zoom out can quickly locate anomalies by avoiding giving each pixel equal attention. In heavy hitter[1] detection for Internet traffic monitoring, traffic flows follow a natural hierarchy based on prefix aggregation of the source or destination IP addresses. Indeed, recent advances in software-defined networking (SDN) allow programmable routers to count aggregated flows that match a given IP prefix [88]. The search space of all traffic flows thus follows a binary tree structure.

Based on the progress flow, the results of this dissertation are partitioned into two parts. In the first part, we propose an optimal nested test plan of which the sample complexity achieves the optimal scaling with the search space. In the second part, with the present of noise in the observations, we develop an information-directed random walk on the tree policy which achieves not only the optimal scaling with the search space but also the optimal scaling with the accuracy requirement. We now introduced the main results in the two parts of the work.

[1]It is a common observation that Internet traffic flows are either "elephants" (heavy hitters) or "mice" (normal flows). A small percentage of high-volume flows account for most of the total traffic [78]. Heavy hitters can be defined as the top flows in terms of weight in total network traffic or flows with a weight exceeding a given threshold.

1.2 Optimal Scaling with the Size of the Search Space

In the first part of this dissertation, we focus on the anomaly detection in the noiseless scenario.

Consider the problem of finding d abnormal processes among n processes where $n \gg d$. The observation from the normal processes and abnormal process are different but fixed (noiseless). Without loss of generality, the observation from the normal processes can be quantized as 0 and from the abnormal processes can be quantized as 1.

Our objective is to develop an algorithm of which the sample complexity is optimal scaling with the size of the search space.

1.2.1 Group Testing Problem

The anomaly detection problem considered here falls into the general class of the group testing problem. Group testing is one of the classical problems that focuses on searching a few rare items among a large group of items. The group testing problem is concerned with identifying defective items in a given population by performing tests over subsets of the population.

Under the classic model, each test gives a binary result, indicating whether the tested group contains any defective items (Boolean test results). The problem was first motivated by the practice of screening draftees with syphilis during World War II, and the idea of testing pooled blood samples from a group of people (rather than testing each person one by one) was initiated by Robert Dorfman [29]. It is not difficult to see that, if the test result indicates no items in the group is defective, then all the items in the group

are cleared in one shot, which saves a large number of tests than testing the items one by one.

A generalization model of the classical Boolean group testing is called the quantitative group testing problem. In a quantitative group testing problem, a test reveals the number of defective items in the tested group, a finer observation model than the binary model assumed in the classic group testing [31]. It is also known as the coin weighing problem with a spring scale first introduced by Shapiro in 1960 [71]. The problem is to identify d counterfeit coins in a collection of n coins. The weights of the authentic and counterfeit coins are known. Thus each weighing gives the number of counterfeit coins in the tested group.

In most of the anomaly detection problems, such as the heavy hitter detection problem and the spectrum sensing, with some prior knowledge of the abnormal items or distribution of the abnormal processes, the decision maker would be able to estimate the number of defective items in the test group. In such problems, a quantitative group testing model is more suitable than a Boolean group testing model. The objective is a test plan with a minimal number of tests identifies all defective items.

1.2.2 Main Results

In Chapeter 2, we consider the quantitative group testing problem under the combinatorial group testing formulation with adaptive test plan for both known and unknown d. This problem with known d was first studied by Aigner and Schughart in [1] in which they established the number of tests required by the optimal nested test plan for identifying d defective items in a population of size n. To our best knowledge, the optimal nested test plan remains open. In this paper, we obtain the optimal nested test plan in

closed form.

The optimal number of tests $N(n,d)$ was given in [1] in the form of inequalities. From these inequalities, we obtain a closed-form expression of $N(n,d)$. We also show that the sequence of $N(n,d)$ in n for fixed d has a clean pattern which can be illustrated in a *frame-segment* structure. However, since $N(n,d)$ is a nonlinear integer-valued function involving multiple layered ceiling functions, directly obtaining the optimal test plan from $N(n,d)$ by solving an integer optimization problem is intractable. Our approach is to first establish three key properties of $N(n,d)$ and of the optimal test plan. Based on these properties, we obtain the optimal test plan in closed form using induction, which also has a clean frame-segment structure corresponding to the pattern of $N(n,d)$. We point out that establishing these properties of $N(n,d)$ itself is nontrivial due to the complex nonlinearity of $N(n,d)$ in both n and d.

We then focus on the application of heavy hitter detection for traffic monitoring and anomaly detection in the Internet and other communication networks. For Internet traffic, it is a common observation that a small percentage of high-volume flows (referred to as heavy hitters) account for most of the total traffic [78]. In particular, it was shown in [37] that the top (in terms of volume) 9% of flows make up 90.7% of the total traffic over the Internet. Quickly identifying the heavy hitters is thus crucial to network stability and security. However, the large number of Internet flows makes individual monitoring extremely inefficient if not impossible. A quantitative group testing approach to heavy hitter detection offers an efficient solution under which the number of required measurements for reliable detection grows logarithmically rather than linearly with the number of flows. Indeed, recent advances in software defined networking (SDN) allow programmable routers to count aggregated flows that match a given IP prefix [88].

The quantitative group testing model stems from the fact that the difference between

the average traffic rates of heavy hitters and normal flows is large, which allows for accurate estimation of the number of heavy hitters from random measurements of the aggregated traffic load. Through simulation examples, we examine the performance of the group testing approach in terms of detection delay, detection accuracy, and counter consumption. Significant improvement over two prevailing sampling-based approach is observed.

Other potential applications include detecting idle channels in the radio spectrum when the signal strength is relatively even across busy channels and much higher than the noise level in idle channels (the high SNR regime).

1.3 Optimal Scaling with the Size of the Search Space and Accuracy in Noisy Scenarios

In the second part of this dissertation, we study the anomaly detection in the noisy scenarios. The objective is to develop a search strategy that minimize the total sample complexity and also meet the reliability requirement.

Consider the problem of finding a few abnormal processes among a large number of processes. Borrowing terminologies from target search, we refer to these processes as cells and the anomalous processes as the targets which can locate in any of all the cells. The observations from sampling a cell are i.i.d. realizations drawn from two different distributions f and g, depending on whether the target is absent or present. The decision maker can take aggregated observations from a chosen subset of processes. The relation between the distribution of the aggregated observation and $\{g_0, f_0\}$ depends on the specific application. The observation models fully specify the noisy observation.

1.3.1 Active Hypothesis Testing

The anomaly detection problem considered here falls into the general class of sequential design of experiments pioneered by Chernoff in 1959 [21] in which he posed a binary (i.e., $M = 2$ for the problem at hand) active hypothesis testing problem. Compared with the classic sequential hypothesis testing pioneered by Wald [82] where the observation model under each hypothesis is predetermined, the sequential design of experiments (a.k.a the active hypothesis testing) has a control aspect that allows the decision maker to choose the experiment to be conducted at each time. Different experiments generate observations from different distributions under each hypothesis. Intuitively, as more observations are gathered, the decision maker becomes more certain about the true hypothesis, which in turn leads to better choices of experiments.

The problem considered here shares similarity with the classic group testing problem (see [30] and references therein). In group testing, the objective is to identify defective items in a large population by performing tests on subsets of items that reveal whether the tested group contains any defective items. Most work on group testing assumes error-free test outcomes. The issue of sample complexity in terms of detection accuracy is absent in the basic formulation.

In this dissertation, we develop an active hypothesis testing plan named as Information-directed Random Walk (IRW) policy, of which the sample complexity is in optimal scaling with the size of the search space as well as the accuracy. The policy also provides a feasible solution for multiple related problems including the noisy group testing, adaptive sampling and channel coding with feedback.

1.3.2 Main Results

In Chapter 3, we consider a large number M of processes, among which L are anomalous. The decision maker aims to search for the anomalous processes by taking (aggregated) observations from a subset of processes, where the chosen subset conforms to a given tree structure. The random observations are i.i.d. over time with a general distribution that may depend on the size of the chosen subset and the number of anomalies in the subset. The objective is a sequential search strategy that adaptively determines which node on the tree to probe at each time and when to terminate the search in order to minimize the sample complexity under a constraint on the error probability.

To fully exploit the hierarchical structure of the search space, the key questions are how many samples to obtain at each level of the tree and when to zoom in or zoom out on the hierarchy. Our approach is to devise an information-directed random walk (IRW) on the hierarchy of the search space. The IRW initiates at the root of the tree and eventually arrives and terminates at the targets (i.e., the anomalous processes) with the required reliability. Each move of the random walk is guided by the test statistic of the sum log likelihood ratio (SLLR) collected from each child of the node currently being visited by the random walk. This local test module ensures that the global random walk is more likely to move toward a target than move away from it and that the walk terminates at a true target with the required probabilistic guarantee on detection accuracy. By constructing a sequence of last passage times of the biased random walk to shrieking subsets of the search space, we show that the sample complexity of the IRW strategy is asymptotically optimal in detection accuracy and logarithmic in M (thus order optimal as determined by the information-theoretic lower bound) provided that the Kullback-Liebler (KL) divergence between aggregated observations in the presence and the absence of anomalous processes are bounded away from zero at all levels of the tree

structure as M approaches infinity. It is thus order optimal in M as determined by the information theoretic lower bound. Using Bernoulli distribution as a case study, we also examine scenarios where higher level observations decay to pure noise as M grows. We establish sufficient conditions on the decaying rate of the quality of the hierarchical observations under which the proposed strategy achieves a sublinear sample complexity in M. This dissertation also includes a detailed discussion on the connection between the active search problem with noisy group testing, adaptive sampling, and channel coding with feedback.

The proposed search strategy is deterministic with search actions explicitly specified at each given time. It involves little online computation beyond calculating the sum log-likelihood ratio and performing simple comparisons. The analysis of its sample complexity in terms of both M and the detection accuracy is based on analyzing a biased random walk on the tree resulted from the search strategy. The desired scaling with M and the detection accuracy is achieved by ensuring that the random walk, initiated at the root of the tree, has a higher probability of moving toward than moving away from the anomalous processes at the leaf level of the tree.

1.4 Organization

The rest of this dissertation is organized as follows. In Chapter 2, we study the optimal nested test plan for the quantitative group testing that achieves the optimal scaling with the size of the search space. In Chapter 3, we study the an information-directed random walk policy that achieves the optimal scaling with both the size of the search space and accuracy. Chapter 4 concludes the dissertation.

CHAPTER 2

ACHIEVING OPTIMAL SCALING WITH THE SIZE OF THE SEARCH SPACE

2.1 Background and Related Work

2.1.1 Classic Group Testing

The group testing problem is concerned with identifying defective items in a given population by performing tests over subsets of the population. The objective is a test plan with a minimal number of tests identifies all defective items.

Under the classic model, each test gives a binary result, indicating whether the tested group contains any defective items. The problem was first motivated by the practice of screening draftees with syphilis during World War II, and the idea of testing pooled blood samples from a group of people (rather than testing each person one by one) was initiated by Robert Dorfman [29].

There are two formulations of the group testing problem, known as *probabilistic group testing* (PGT) and *combinatorial group testing* (CGT). The former is a Bayesian formulation that assumes a probabilistic model on the defective items and aims to minimize the expected number of tests for identifying all defective items [74]. The latter is a minimax formulation that assumes a deterministic value d for the total number of defective items and aims to minimize the number of tests in the worst case (among all compositions of the defective set of size d) [31, 55, 62].

Under both formulations, the test plans can be adaptive or non-adaptive. Adaptive

test plans are sequential in nature: which group to test next depends on the outcome of the previous tests. The studies in [55, 62, 74] mentioned above all focus on adaptive test plans. Non-adaptive group testing is a one-stage problem in which all actions are determined before any test is performed. Non-adaptive test plans are often represented by matrices [32, 68].

The classic group testing problem has seen a wide range of applications, including multiaccess communications [6, 84, 85], idle channel detection in the radio spectrum [72], compressed sensing [17], network tomography [18], and anomaly detection [58, 77]. In particular, non-adaptive group testing has been widely applied to DNA sequencing and DNA library screening [5, 68].

2.1.2 Quantitative Group Testing

In a quantitative group testing problem, a test reveals the number of defective items in the tested group, a finer observation model than the binary model assumed in the classic group testing [31]. It is also known as the coin weighing problem with a spring scale first introduced by Shapiro in 1960 [71]. The problem is to identify d counterfeit coins in a collection of n coins. The weights of the authentic and counterfeit coins are known. Thus each weighing gives the number of counterfeit coins in the tested group.

Most studies on quantitative group testing focus on non-adaptive test plans, see, for example, [35, 38, 63] on the case of unknown d and [23, 28] on the case of known d. Adaptive test plans have been studied mostly for the special case of $d = 2$ (see [2, 39, 46]). Only a couple of results are available on adaptive test plans for the general case of $0 < d < n$. In particular, Aigner and Schughart considered a class of adaptive test plans with a *nested* structure [1]. Specifically, in a nested test plan, once a test reveals

a group containing defective items, the next test must be a proper subset of this group. They established the performance (i.e., the number of required tests) of the optimal nested test plan. The optimal nested test plan itself, however, was not obtained. In [10], Bshouty developed a semi-adaptive test plan that integrates a bisecting search with a non-adaptive test plan. It was shown that this semi-adaptive test plan can be constructed in polynomial time and has a performance no worse than twice of the information-theoretic lower bound. However, the algorithm may fail to construct a valid test plan in certain cases[1].

The applications of quantitative group testing include the uniquely decodable codes for the noiseless n-user adder channel problem [16], and the construction of unknown graphs from additive queries [2, 23]. Several variations of the problem can be found in [11, 43, 45].

2.1.3 Related Work

Much of the related work has been discussed in the preceding chapters. Here we provide additional related work, focusing on the comparison between adaptive and non-adaptive group testing approaches and the connection between quantitative group testing and compressed sensing.

Adaptive vs. Non-Adaptive Group Testing

Most work on group testing focuses on non-adaptive test plans. A non-adaptive test plan can be represented by a binary measurement matrix with columns corresponding

[1]One such example is when $n = 200$ and $d = 52$, the algorithm fails to construct the corresponding $(3, 3, 4, 5, 6, 7, 8, 8, 9, 9)$-*Detection Matrix*.

to items, rows corresponding to tests, and the (i, j)th element indicating whether item j is included in the ith group test. Constructing the measurement matrix can be cast as a source coding problem, and the superimposed code and the uniquely decipherable code have been used in developing non-adaptive test plans (see, for example, [7, 33, 34, 42, 56]). It is this connection to source coding that brings mathematical tractability to non-adaptive group testing, a treat seldom enjoyed by adaptive group testing. Allowing parallel implementation with all tests run simultaneously also makes non-adaptive test plans attractive in applications that involve a lengthy delay in obtaining test results. The disadvantages of non-adaptive test plans lie in the computational complexity of the coding/decoding processes, high storage requirement, and difficulty to adjust to cases with unknown or time-varying population compositions $\{n, d\}$.

Adaptive test plans, in contrast, are more suitable for online applications where the values of n and/or d are not prefixed. Furthermore, the optimal nested test plan developed in this work is given in closed form and has a clean frame-segment structure; little offline or online computation is needed. The inherent tree structure of the nested test plan also leads to low memory requirement. It is thus particularly attractive for online applications such as real-time heavy hitter detection where n and d are not prefixed and computational, memory, and counter resources are stringent.

Connection with Compressed Sensing

The quantitative group testing problem shares similarity with the compressed sensing problem. In compressed sensing, the objective is to recover a sparse signal from linear measurements. Specifically, given an n-dimensional sparse signal with a support size d, the goal is to identify the support set (non-zero elements of the signal) with a minimum number of projections. The differences between compressed sensing and quantitative

group testing are in the signal model and constraints on the measurement/projection matrix. Most work on compressed sensing assumes real-valued signals and allow real-valued measurement matrices. Quantitative group testing, when viewed as compressed sensing, deals with binary signals that are not necessarily sparse and require the measurement matrix to be binary valued. There are a number of non-adaptive compressed sensing algorithms in the literature that result in binary-valued measurement matrices (see, for example, [8, 19, 50, 52, 86, 87]). However, in addition to the sparsity requirement, these non-adaptive strategies suffer the same difficulties as non-adaptive group testing algorithms in online applications as discussed above.

Several adaptive compressed sensing algorithms exist in the literature [47,51,53,65]. They were shown to outperform non-adaptive algorithms in sample complexity and detection performance. Most of the adaptive compressed sensing algorithms, however, are not directly applicable to quantitative group testing due to the real-valued measurement matrices. The only exceptions are [51, 65], in which two similar bisecting search approaches were introduced. While the problems formulated in [51, 65] were to estimate a real-valued sparse signal under certain constraints, the bisecting search approach proposed there can be applied to the quantitative group testing problem and constitutes a suboptimal nested test plan. In this work, we develop the optimal nested test plan for combinatorial quantitative group testing.

2.2 Problem Formulation

Consider a population of n items. It is known that among these n items, d are defective (the issue of unknown d is addressed in Section 2.3.3). Let (n, d) denote the corresponding quantitative CGT problem. We assume that $1 \leq d \leq n-1$ to avoid the trivial

scenarios of $d = 0$ and $d = n$.

For a given (n, d), an adaptive test plan π is a sequence of decision rules $\{\pi_1, \pi_2, \ldots\}$ where π_t maps from the outcomes of the previous $t-1$ tests to the subset of items to be tested in the tth test. With a slight abuse of notation, π_t is also used to denote the subset of items tested in the tth test under test plan π. Let $N_\pi(n, d; \mathcal{D})$ denote the number of tests required for identifying all d defective items under π when the d defective items are specified by the set $\mathcal{D}$. Note that n and d are known while $\mathcal{D}$ is unknown and is what the test plan needs to identify. Under the combinatorial formulation, the performance of a test plan is determined by the worst instance of $\mathcal{D}$ among all subsets with size d. The performance of π, denoted by $N_\pi(n, d)$, is thus given by

$$N_\pi(n, d) = \max_{\mathcal{D} \subset [n], |\mathcal{D}| = d} N_\pi(n, d; \mathcal{D}), \tag{2.1}$$

where $[n]$ denotes the set of all n items.

In this work, we focus on a family of test plans that exhibit a tree structure. This family is referred to as the nested test plan as defined below.

Definition 1. *An adaptive test plan $\pi = \{\pi_1, \pi_2, \ldots\}$ is a nested test plan if for all $t \geq 1$ and $k \geq 1$, the tested groups π_t and π_{t+k} at the tth and the $(t+k)$th tests satisfy either $\pi_t \cap \pi_{t+k} = \pi_{t+k}$ or $\pi_t \cap \pi_{t+k} = \emptyset$.*

Based on the above definition, it is not difficult to show that for every instance of $(n, d; \mathcal{D})$, the $N_\pi(n, d; \mathcal{D})$ tested groups $\{\pi_1, \pi_2, \ldots, \pi_{N_\pi(n,d;\mathcal{D})}\}$ form a tree. Specifically, consider a graph with $N_\pi(n, d; \mathcal{D})$ nodes representing each of the tested groups and a root node representing the set $[n]$ of the entire population. An edge exists between two nodes if and only if one of them is the smallest superset of the other. It can be shown such a graph resulting from a nested test plan is acyclic.

Our objective is an optimal nested test plan π^* given by

$$\pi^* = \arg\min_{\pi \in \Pi} N_\pi(n, d), \tag{2.2}$$

where Π denotes the family of all nested test plans. To simplify the notation, the performance of the optimal nested test plan π^* is denoted by $N(n, d)$ (rather than $N_{\pi^*}(n, d)$).

For a given CGT (n, d), due to the symmetry among items, the worst-case performance $N_\pi(n, d)$ of any test plan π depends on the first test only through the size of the tested group but not the specific composition of the group. Suppose that the first test consists of m items and the outcome reveals that d_1 items among these m are defective. For a nested test plan, this first test decomposes the original CGT problem of (n, d) into two independent CGT problems of (m, d_1) and $(n - m, d - d_1)$. Obviously, d_1 cannot exceeds m or d. At the same time, d_1 cannot be smaller than 0 or $d - (n - m)$ (the latter is due to the fact that the $n - m$ untested items consist of at most $(n - m)$ defective items). Combined with the minimax nature of the CGT formulation, this leads to the following recursive equation for $N(n, d)$:

$$N(n, d) = 1 + \min_m \max_{d_1} \{N(m, d_1) + N(n - m, d - d_1)\}, \tag{2.3}$$

where the maximization over d_1 is among integers in the range of $\max\{0, d + m - n\}$ to $\min\{m, d\}$ and the minimization over m can be set to integers $1, 2, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$ (since testing a group of size m is equivalent to testing a group of size $n - m$).

Due to this decomposition of the problem into two independent problems of smaller sizes, a nested test plan is fully specified once the first test is determined for all possible population sizes n and all possible numbers d of defective items. Furthermore, since the composition of the tested group is inconsequential, specifying the size m of the first group test for all n and d suffices. Let $M(n, d)$ denote the value of m that achieves $N(n, d)$ in (2.3), i.e.,

$$M(n, d) = \min_m \max_{d_1} \{N(m, d_1) + N(n - m, d - d_1)\}. \tag{2.4}$$

The values of $M(n,d)$ for all $n \geq 1$ and $0 \leq d \leq n$ specify the optimal nested test plan for all CGT problems. We point out that when there are multiple values of the group size m that achieve the minimum in (2.4), $M(n,d)$ is set to the smallest such value. A smaller group size is often preferred in practical applications.

The focus on nested test plan is motivated by its analytical tractability, its simple implementation, and its order optimality. Without imposing any structure, the optimal test plan is analytically intractable in general. Obtaining the optimal test plan numerically through exhaustive search is computationally prohibitive due to the combinatorial nature of the problem. The nested structure, however, leads to the clean recursive formulas in (2.3, 2.4), offering the possibility of explicit analytical characterizations. Nested test plans also enjoy simpler implementation due to the tree-structured splitting of previously tested groups. This tree structure results in lower memory requirement for storing all past test outcomes. It also allows maintaining a certain contiguous property in each tested group, which is often desirable in practice. For example, for the application of heavy hitter detection, the contiguous property is in terms of all flows in the tested group sharing a common IP prefix, which simplifies the router configuration for packet count of the aggregated flow. For the application of spectrum sensing, the contiguous property is in terms of adjacency in the spectrum, which eases filter implementation. Lastly, the optimal nested test plan is order optimal among all test plans as shown in Section 2.3.4.

2.3 The Optimal Nested Test Plan

In this section, we establish the optimal nested test plan in closed-form. This result hinges on a compact closed-form expression of $N(n,d)$ and its geometric block-constant structure as established in Lemma 1 below.

2.3.1 $N(n, d)$ and Its Geometric Block-Constant Structure

In QGT, a test outcome reveals the number of defective items, thus also the number of non-defective items. This symmetry between defective and non-defective items readily leads to $N(n, d) = N(n, n-d)$. It thus suffices to assume $d \leq \left\lfloor \frac{n}{2} \right\rfloor$ unless otherwise noted.

The performance of the optimal nested tested plan is given in the following lemma.

Lemma 1. *For a CGT problem (n, d) with $d \leq \frac{n}{2}$, we have*

$$N(n, d) = (l+1)d + k - 1, \tag{2.5}$$

where

$$l = \lceil \log_2 (n/d) \rceil - 1, \tag{2.6}$$

$$k = \lceil n/2^l \rceil - d. \tag{2.7}$$

Proof. The proof is based on the characterization of $N(n, d)$ given in [1] in the form of the following three inequalities.

$$N(2d, d) \geq 2d - 1, \tag{2.8}$$

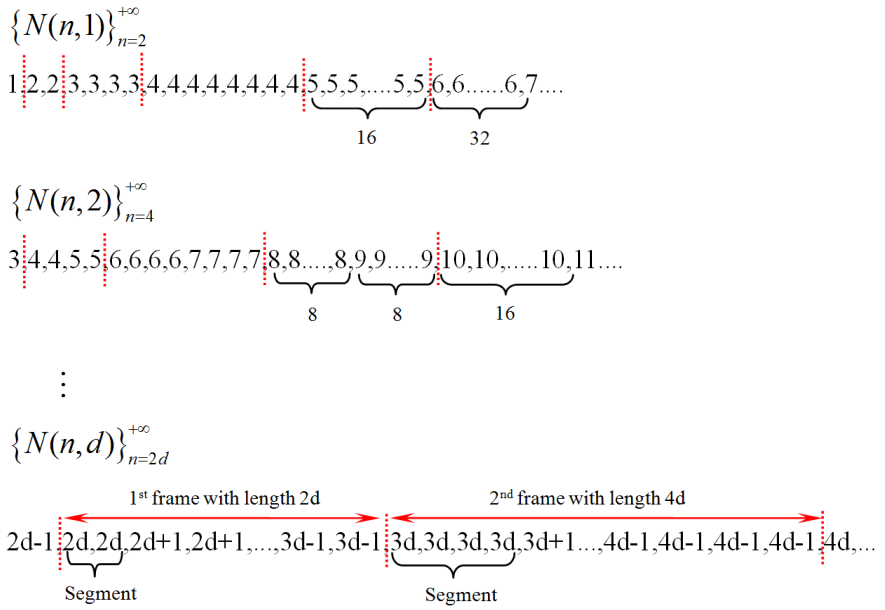
$$N((d+i)2^{t-1}, d) \leq td + i - 1, \tag{2.9}$$

$$N((d+i)2^{t-1} + 1, d) \geq td + i, \tag{2.10}$$

where $t \geq 2,\ d \geq 1,\ 0 \leq i \leq d-1$. Detailed of the proof are given in Appendix A.1. □

Lemma 1 reveals an interleaved block-constant structure with geometrically growing block length of rate 2. As illustrated in Table 2.1, the sequence of $N(n, d)$ in terms of n for a fixed d consists of *frames*, with each frame containing d *segments*. The two positive integers l and k given in (2.6) and (2.7) are, respectively, the frame index and the segment

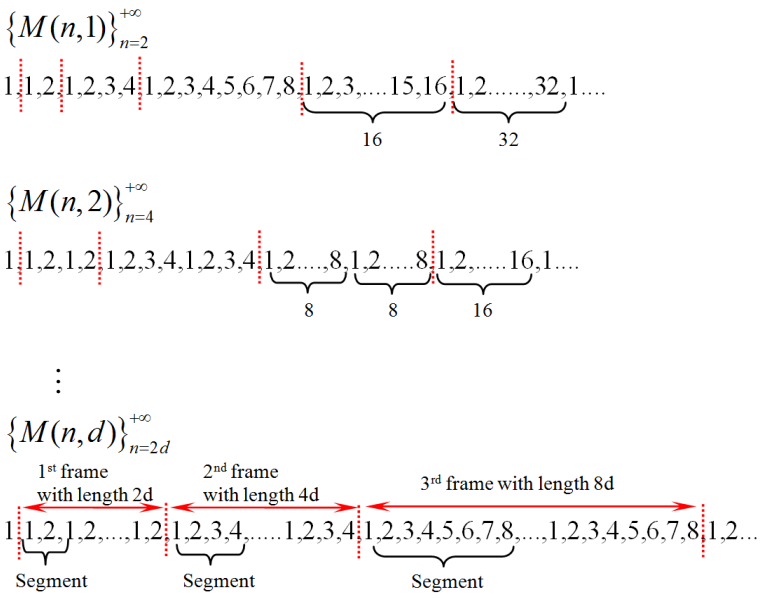
Table 2.1: The Frame-Segment Structure of $N(n,d)$



index. Specifically, each sequence $N(n,d)$ starts at $n = 2d$ with $N(2d,d) = 2d-1$ (recall that it is sufficient to consider $d \leq \left\lfloor \frac{n}{2} \right\rfloor$). Following this initial value, the rest of the sequence is partitioned into frames with the frame length doubled from one frame to the next. Each frame consists of d segments of equal length with a segment length of 2^l in the lth frame ($l = 1, 2, \ldots$). The value of $N(n,d)$ is the same within a segment and increases by 1 from one segment to the next.

We point out that while $N(n,d)$ was determined in [1], it was specified through the three inequalities given in (2.8-2.10). The expression given in Lemma 1 is not only more compact but also reveals the frame-segment structure of $N(n,d)$. As shown in Section 2.4, this frame-segment structure of $N(n,d)$ is the key to establishing the optimal nested test plan.

Table 2.2: The Frame-Segment Structure of $M(n,d)$



2.3.2 The Optimal Nested Test Plan

The theorem below characterizes the optimal nested test plan $M(n,d)$ in closed form for all n and d.

Theorem 1. *For a CGT problem* (n,d) *with* $d \leq \frac{n}{2}$*, we have*

$$M(n,d) = n - 2^l(d+k-1), \tag{2.11}$$

where l *and* k *are the frame and segment indexes as given in* (2.6) *and* (2.7)*. For* $d > \frac{n}{2}$*, we have*

$$M(n,d) = M(n,n-d). \tag{2.12}$$

The theorem above fully specifies the optimal nested test plan. A pseudo code implementation with a recursively called subroutine is given below.

Algorithm 1: Optimal Nested Test Plan

Input: $[n]$: a group of n items;

d: number of defectives in the group.

Output: $\mathcal{D}$: the set of defective items.

procedure TEST($[x]$)

    return The number of defective items in set $[x]$.

end procedure

procedure NESTED($[n], d$)

    Initialize: $\mathcal{D} = \emptyset$

    if $d = 0$ **then**

        return $\mathcal{D} = \emptyset$

    else if $d = n$ **then**

        return $\mathcal{D} = [n]$

    else

        $[M(n, d)]$ = a subset of $[n]$ with size $M(n, d)$

        d_1 = TEST($[M(n, d)]$)

        $\mathcal{D} = \mathcal{D} \cup$ NESTED($[M(n, d)], d_1$)

        $\mathcal{D} = \mathcal{D} \cup$ NESTED($[n] \setminus [M(n, d)], d - d_1$)

        return $\mathcal{D}$

    end if

end procedure

$M(n,d)$ as a sequence of n for a fixed d has the same frame-segment structure as $N(n,d)$. Specifically, each sequence starts at $n = 2d$ with $M(2d,d) = 1$. The values of $M(n,d)$ in each segment of the lth frame are consecutive integers from 1 to 2^l.

We postpone the proof of Theorem 1 to Section 2.4 where we establish several key properties of $N(n,d)$ that will be used in the proof.

2.3.3 The Optimal Nested Test Plan for CGT with Unknown d

We have so far focused on the standard CGT formulation which assumes a prior knowledge on the total number of defective items in the given population. For applications where this prior knowledge is unavailable, the question is how to start the first test: for any population size n, should the first test be carried over the entire population or a proper subset of the population with the size potentially depending on n? The answer is given in the following theorem.

Theorem 2. *For a CGT problem with a population size n and an unknown number of defective items, the optimal nested test plan first tests the entire population.*

Proof Let d denote the number of defective items in the population of n. Suppose that the first test is not carried over the entire population, but rather on a subset of n_1 items. Due to the nested structure, any nested test plan π will break the problem with an unknown d into a sequence of CGT problems (n_k, d_k) $(k = 1, 2, \ldots, K)$ for some integers $K > 0$, $\{n_k\}_{k=1}^K$ with $\sum_k n_k = n$, and $\{d_k\}_{k=1}^K$ with $\sum_k d_k = d$. Specifically, the test plan first tests a group of size n_1, and with one test revealing the number d_1 of defective items in this group, the test plan then resolves the CGT problem (n_1, d_1). Subsequently, the test plan determines the size n_2 of the next group of unidentified items to test, where the

choice of n_2 may depend on the outcomes of past tests. The procedure continues until all items are identified. We thus have

$$N_\pi(n) = K + \sum_{k=1}^{K} N_\pi(n_k, d_k).$$

Now consider the CGT problem (n, d). A slight modification of π that omits the group test of the last set of n_K unidentified items (since the number of defective items in this last set can be deduced from past tests when d is known) gives a valid nested test plan for the CGT problem (n, d). We thus have

$$N_\pi(n) \geq N(n, d) + 1.$$

We then arrive at Theorem 2 by noticing that the lower bound of $N(n, d) + 1$ can be achieved by first testing the entire population and that π is an arbitrary nested test plan.

With the first test revealing the total number d of defective items, the problem is then reduced to a CGT of (n, d).

2.3.4 Order Optimality and the Approximation Ratio of the Optimal Nested Test Plan

The logarithmic order of $N(n, d)$ in terms of n can be readily seen from the closed-form expression. Specifically, we can write $N(n, d)$ in (2.5) as

$$N(n, d) = \left\lceil \log_2 \frac{n}{d} \right\rceil \cdot d + \left\lceil \frac{n}{2^l} \right\rceil - d - 1, \tag{2.13}$$

where $\left\lceil \frac{n}{2^l} \right\rceil - d - 1$ is bounded between 0 and $d - 1$. We compare below the order of $N(n, d)$ with that of $N^*(n, d)$, the minimum number of tests achievable among all test plans.

Likening the group testing problem (n, d) to a source coding problem with the entropy of the source given by $\log_d \binom{n}{d}$ and each test outcome representing one letter in the corresponding codeword, we can easily obtain a lower bound of $\log_d \binom{n}{d}$ (the minimum expected codeword length) on $N^*(n, d)$. Thus, for all fixed d, the optimal nested test plan has a constant (i.e., independent of n) approximation ratio that is asymptotically bounded by

$$\lim_{n\to\infty} \frac{N(n, d)}{N^*(n, d)} \leq \log_2 d. \tag{2.14}$$

In other words, for all fixed d, the optimal nested test plan is order optimal among all test plans.

Note that whether the information-theoretic lower bound of $\log_d \binom{n}{d}$ is achievable is still an open question, since not every coding scheme can be mapped to a valid test plan. In particular, while a source code has no constraint in choosing each letter of a codeword, the sequence of test outcomes are bound by the specific configuration of the given population. For example, a test outcome cannot take a value greater than the size of the tested group, and the test outcome of a subset of a previously tested group must be consistent with the test outcome of that group. In fact, a negative answer has been established when we restrict to non-adaptive test plans. Thus, the asymptotic bound on the approximation ratio given in (2.14) may be a pessimistic one. A more detailed discussion on achievable performance and a comparison between quantitative and Boolean group testing are given in Section 2.5.

2.4 Properties of $N(n, d)$ and Proof of Theorem 1

2.4.1 Properties of $N(n, d)$

We first establish three properties of $N(n, d)$, which will be used in proving the closed form of $M(n, d)$ in Theorem 1.

Properties:

[P1] $\{N(n, d)\}_{d=0}^{\lfloor n/2 \rfloor}$ is a strictly increasing sequence in d, i.e.,

$$N(n, d) > N(n, d-1), \quad \forall 1 \le d \le \left\lfloor \frac{n}{2} \right\rfloor.$$

[P2] $\{N(n, d)\}_{d=0}^{n}$ is a concave sequence in d, i.e., for all $1 \le d \le n-1$, we have

$$N(n, d+1) - N(n, d) \le N(n, d) - N(n, d-1).$$

[P3] For all $d \le \left\lfloor \frac{n}{2} \right\rfloor$ and $m \le \left\lfloor \frac{n}{2} \right\rfloor$, if

$$N(m, 0) + N(n-m, d) \ge N(m, 1) + N(n-m, d-1),$$

then for all $d_1 = 1, 2, \ldots, \min\{m,\ d\}$,

$$N(m, 0) + N(n-m, d) \ge N(m, d_1) + N(n-m, d-d_1).$$

The strict increasing property [P1] is proved via induction in n and is the key property used to prove [P2]. [P3] is proved based on [P2] and is the main tool for proving Theorem 1. It is used to show that, when $m = M(n, d)$, the worst case occurs at $d_1 = 0$, i.e., the maximization over d_1 in (2.4) is achieved at $d_1 = 0$. The proof of these three properties can be found in Appendix A.2.

2.4.2 Proof of Theorem 1

We now provide a proof of Theorem 1. It suffices to consider $d \leq \frac{n}{2}$. The proof hinges on [P3] shows that when $m = M(n, d)$ the worst case occurs at $d_1 = 0$. Therefore, $N(n, d)$ equals $1 + N(n - M(n, d), d)$, which is the number of tests in the previous segment plus 1 according to the frame-segment structure of $N(n, d)$. The detailed proof follows below.

We first establish the initial value $M(2d, d) = 1$ of every sequence d. From Lemma 1, we have $N(2d, d) = 2d - 1$, which can be achieved by testing all but the last item one by one, i.e., $M(2d, d) = 1$.

For $n > 2d$, recall the frame-segment of $N(n, d)$ as illustrated in TABLE 2.2. Consider the x-th ($x = 1, \ldots, 2^l$) element in the k-th segment of the l-th frame, i.e.,

$$n = 2^l(d + k - 1) + x.$$

Then (2.11) is equivalent to

$$M(2^l(d + k - 1) + x, d) = x. \tag{2.15}$$

For notational simplicity, when the test plan selects the subset with size m to test, let $\phi(m;\, n, d)$ denote the worst case number of tests for the subsequent testing under the optimal nested test plan, i.e.,

$$\phi(m;\, n, d) = \max_{d_1} \{N(m, d_1) + N(n - m, d - d_1)\}. \tag{2.16}$$

Recall that $M(n, d)$ is chosen as the minimum value of the group size m that achieves the optimal performance $N(n, d)$. To show (2.15), it suffices to show that

$$1 + \phi(m;\, n, d) \begin{cases} > N(n, d) & \text{when } m < x, \\ = N(n, d) & \text{when } m = x. \end{cases} \tag{2.17}$$

When $m < x$, we have

$$\begin{aligned}
&1 + \phi(m;\, 2^l(d+k-1)+x, d) \\
&\overset{(a)}{\geq} 1 + N(m,0) + N(2^l(d+k-1)+x-m, d) \\
&\overset{(b)}{>} N(2^l(d+k-1)+x, d),
\end{aligned}$$

where (a) holds by setting $d_1 = 0$ in (2.16) and (b) follows from the fact that $N(2^l(d+k-1)+x-m, d) = N(2^l(d+k-1)+x, d)$ since they are in the same segment.

When $m = x$, based on Lemma 1, we have

$$\begin{aligned}
&N(x,0) + N(2^l(d+k-1), d) - N(2^l(d+k-1), d-1) - N(x,1) \\
&= (l+1)d + k - 2 - (l+1)(d-1) - k + 1 - N(x,1) \\
&= l - N(x,1) \geq 0,
\end{aligned}$$

i.e.,

$$N(x,0) + N(2^l(d+k-1), d) \geq N(x,1) + N(2^l(d+k-1), d-1). \tag{2.18}$$

With (2.18), based on [P3], we thus have

$$\begin{aligned}
&1 + \phi(x;\, 2^l(d+k-1)+x, d) \\
&= 1 + \max_{d_1}\{N(x,d_1) + N(2^l(d+k-1), d-d_1)\} \\
&= 1 + N(x,0) + N(2^l(d+k-1), d) \\
&= (l+1)d + k - 1 \\
&= N(2^l(d+k-1)+x, d),
\end{aligned}$$

i.e., $m = x$ achieves the optimal performance $N(2^l(d+k-1)+x, d)$. We then conclude that $M(2^l(d+k-1)+x, d) = x$.

Table 2.3: A comparative summary of boolean and quantitative CGT results.

		Lower Bound	Upper Bound
Non-adaptive	Boolean	$\frac{d^2 \log_2 n}{24 \log_2 d}$ [31]	$\frac{4d^2 \log_2^2 n}{\log_2^2(d \log_2 n)}$ [57]
	Quantitative	$2d \log_d \left(\frac{n}{d}\right)$ [28, 63]	$4d \log_d \left(\frac{n}{d}\right)$ [44] (Non-constructive)
Adaptive	Boolean	$\log_2 \binom{n}{d}$	$\log_2 \binom{n}{d} + d$ [31]
	Quantitative	$\log_d \binom{n}{d}$	$\left\lceil \log_2 \frac{n}{d} \right\rceil \cdot d + \left\lceil \frac{n}{2^l} \right\rceil - d - 1$ [this work]

2.5 Comparison between Quantitative and Boolean Group Testing

It is informative to summarize and compare the best known results for quantitative and Boolean CGT. In particular, it is of interest to examine the potential gain offered by quantitative test outcomes over Boolean test outcomes.

2.5.1 Comparison for Cases with Known d

We first consider the case when the total number d of defective items is known. We summarize in TABLE 2.3 the best known lower bounds and upper bounds for Boolean CGT and quantitative CGT.

For Boolean CGT, when restricted to non-adaptive test plans, the tightest lower bound on the number of required tests was established in [31] to be $\frac{d^2 \log_2 n}{24 \log_2 d}$, which is strictly greater than the information-theoretic lower bound of $\log_2 \binom{n}{d}$. In other words, the information-theoretic lower bound cannot be achieved by non-adaptive test plans. The best known non-adaptive test plan appears to be the one developed in [57] based on disjunct code. However, there remains a gap between the performance of this best known test plan and the tightest lower bound (see Table III). This gap has also been

studied in [3] from the perspective of the *asymptotic rate* of the Boolean group testing algorithms. When considering adaptive test plans, the information-theoretic lower bound can be asymptotically achieved by the adaptive Generalized Binary Splitting (G-BS) algorithm developed in [31] for all fixed d.

For quantitative CGT, the tightest lower bound of the non-adaptive test plan on the number of required tests is $2d \log_d \left(\frac{n}{d}\right)$ [28, 63], which is about twice the information-theoretic lower bound of $\log_d \binom{n}{d}$ for large n. Grebinski and Kucherov [44] established the *existence* of a non-adaptive test plan with a performance of $4d \log_d \left(\frac{n}{d}\right)$. However, this upper bound result is non-constructive, and no non-adaptive test plan is known to achieve this upper bound. For the adaptive test plans, the optimal nested test plan established in this work appears to be the first for the general quantitative CGT problem and achieves order optimality for all fixed d.

The comparison in TABLE 2.3 shows that results on quantitative CGT are much less complete than Boolean CGT. In particular, it remains to be an open question whether the information-theoretic lower bound can be achieved by an adaptive test plan for quantitative CGT. Consequently, whether a gain of $\log_2 d$ indicated by the information-theoretic lower bound for quantitative CGT over Boolean CGT can be realized remains elusive. Nonetheless, it can be shown that the worst-case performance of the optimal nested test plan is strictly better than the Boolean CGT lower bound $\log_2 \binom{n}{d}$. In Fig. 2.1, we compare the average performance of the optimal nested test plan with quantitative test outcomes with that of GBS, the best known adaptive test plan for Boolean CGT. The objective is to illustrate the potential gain offered by quantitative test outcomes over Boolean test outcomes. As shown in Fig. 2.1, for a CGT with $n = 500$, the gain increases with d and can be up to 25%.

2.5.2 Comparison for Cases with Unknown d

We now consider the case with unknown d. For quantitative CGT, a single test of the entire population reveals d and reduces the problem to a CGT (n, d) with a known d. For Boolean CGT, however, most existing test plans rely on the knowledge of d and do not easily extend to the case with unknown d. For example, the aforementioned best known non-adaptive test plan and the best adaptive test plan GBS both require the knowledge of d. How to estimate d based on Boolean test outcomes is highly nontrivial.

One approach to Boolean CGT with unknown d is binary splitting, which is also asymptotically optimal. In Fig. 2.2, we compare the average performance of the nested test plan with quantitative test outcomes with that of a bisection search for Boolean CGT. Fig. 2.2 shows that when d is unknown, the gain offered by quantitative test outcomes over Boolean test outcomes increases, with up to 50% gain for the same CGT problem tested in Fig. 2.1.

2.6 Application to Heavy Hitter Detection

In this section, we study the application of quantitative group testing to the heavy hitter detection problem.

Consider a network consisting of n flows, each modeled as a random process with a certain packet arrival rate. Assume that among the n flows, n_x are heavy hitters with rate λ_x, and $n - n_x$ are normal flows with rate λ_y. Define

$$\rho = \frac{n_x}{n}, \tag{2.19}$$

$$\eta = \frac{n_x \lambda_x}{n_x \lambda_x + (n - n_x)\lambda_y} \tag{2.20}$$

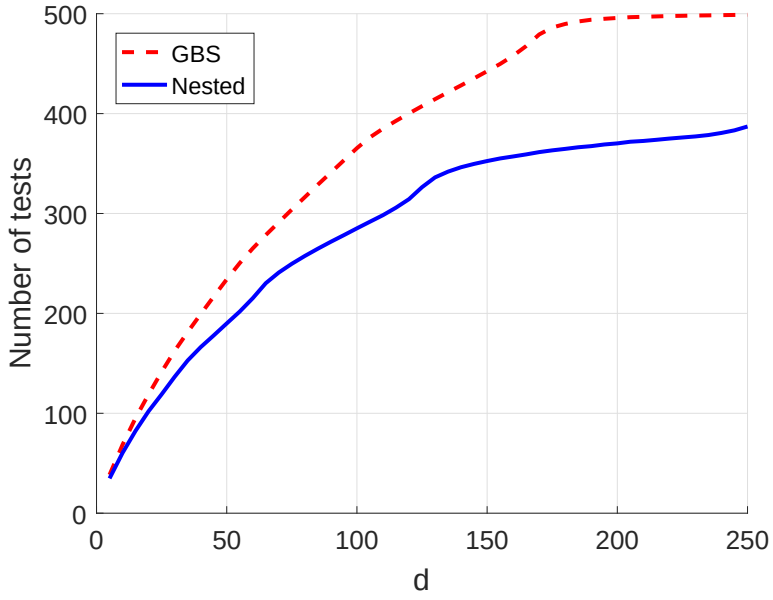


Figure 2.1: Comparison of the optimal nested test plan to the generalized binary splitting (GBS) test plan with known d ($n = 500$, 1000 Monte Carlo runs).

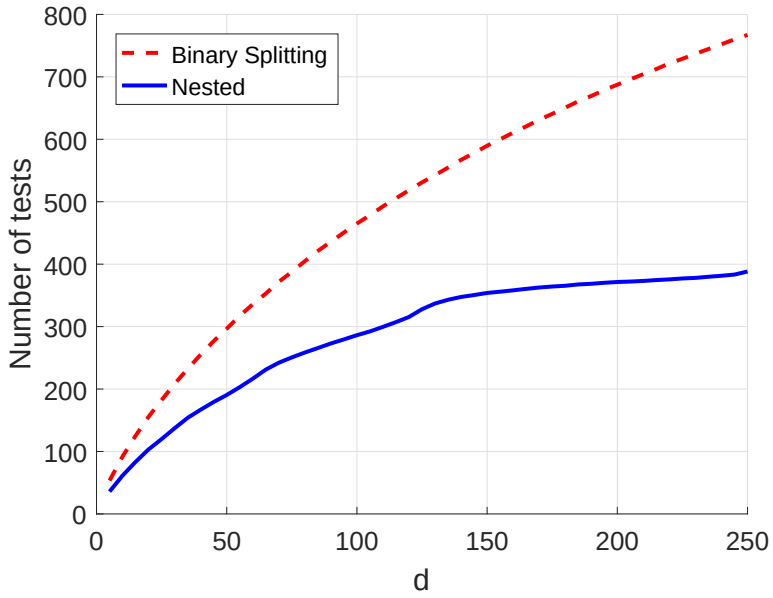


Figure 2.2: Comparison of the optimal nested test plan to the binary splitting test plan with unknown d ($n = 500$, 1000 Monte Carlo runs).

as the fraction of heavy hitters in terms of the number of flows and the total traffic volume, respectively. For Internet traffic, we typically have ρ around 10% to 20% and η around 80% to 90%.

The problem is to identify the n_x heavy hitters quickly and reliably. The performance metrics of interest are detection delay, detection accuracy, and counter consumption. Detection delay is defined as the average time taken to identify all heavy hitters. Detection accuracy is measured by the false positive rate α and false negative rate β defined as

$$\alpha = \frac{\text{Number of falsely identified heavy hitters}}{n - n_x}, \tag{2.21}$$

$$\beta = \frac{\text{Number of missed heavy hitters}}{n_x}. \tag{2.22}$$

Counter consumption is given by the number of flow counters required by a heavy hitter detector. In the group testing algorithm, each test requires a counter, and the counter can be reused. Since flow counters rely on the high-speed TCAM (ternary content-addressable memory) entries which are scarce resources in routers, detectors with low counter consumption are desired.

Without loss of generality, the arrival rate λ_y of normal flows in all simulation examples is normalized to 1. The time unit is thus determined by the expected inter-arrival time of a normal flow, which is in the millisecond scale or smaller in typical Internet traffic.

2.6.1 Quantitative Group Testing for Heavy Hitter Detection

In the quantitative group testing formulation, it is assumed that the test result reveals the number of defective items without any error. A test plan can thus correctly identify all defective items. In the application of heavy hitter detection, the number of heavy hitters

needs to be estimated from random observations of packet arrivals in an aggregated flow. The estimation errors lead to false positives and false negatives in the final detection result. We show below via simulation examples that the large gap in the arrival rates of normal flows and heavy hitters allow accurate estimation of the number of heavy hitters from random packet arrivals. Consequently, the optimal nested test plan given in Theorem 1 offers attractive performance in detection accuracy.

In the first example, we assume that each flow is an independent Possion process. We employ the maximum likelihood estimator (MLE) in estimating the number of heavy hitters in each group test. Consider, without loss of generality, the first group test that aggregates all n flows. Let z denote the number of packet arrivals observed in T time units in the aggregated flow. It is easy to see that the likelihood function is given by

$$L(n_x|z) = z\log[(n\lambda_y + n_x(\lambda_x - \lambda_y))T] - (n\lambda_y + n_x(\lambda_x - \lambda_y))T - \log(z!).$$

The ML estimate of n_x is given by

$$\hat{n}_x = \arg\max_{n_x=0,1,\ldots,n} L(n_x|z). \tag{2.23}$$

The above integer optimization can be simplified to the following

$$\hat{n}_x = \arg\max_{n_x=i_0,i_0+1} L(n_x|z), \tag{2.24}$$

where $i_0 = \left\lfloor \frac{(z/T)-n\lambda_y}{\lambda_x-\lambda_y} \right\rfloor$. The above simplification results from the fact that $L(n_x|z)$, when viewed as a function of a real-valued argument n_x, is unimodal with the maximum value achieved at $\frac{(z/T)-n\lambda_y}{\lambda_x-\lambda_y}$.

From Fig. 2.3 we observe that for all typical values of ρ and η, the group testing approach offers good detection reliability using only $T = 2$ time units for each group test. Furthermore, the detection performance improves when η increases and/or ρ decreases, since both result in a larger gap between λ_x and λ_y, thus better estimates of the number of heavy hitters from random packet arrivals.

The observation that a larger gap between the rates of heavy hitters and normal flows leads to better detection accuracy may also be deduced from the Cramér-Rao lower bound on the mean-squared error (MSE) of estimating n_x. Treating n_x as a real-valued argument, we obtain the lower bound as

$$\mathrm{Var}(\hat{n}_x) \geq \frac{n\lambda_y + (\lambda_x - \lambda_y)n_x}{T(\lambda_x - \lambda_y)^2}, \tag{2.25}$$

showing smaller estimation error when $(\lambda_x - \lambda_y)$ increases for a fixed λ_y. Since the likelihood function is unimodal, we may expect that the MSE in estimating a real-valued proxy of n_x preserves the general property of the original integer estimation problem.

The MLE requires the knowledge of the flow distribution and can be computationally expensive for general distributions. An alternative is a simple sample mean estimator (SME) given by

$$\hat{n}_x = \left[\frac{z/T - n\lambda_y}{\lambda_x - \lambda_y}\right], \tag{2.26}$$

where $[\cdot]$ denotes the operation of taking the nearest integer.

The detection performance of the optimal nested test plan with SME for log-normal distributed flows is shown in Fig. 2.4. By increasing the observation time to $T = 5$ for each group test, SME leads to similar detection accuracy for heavy-tailed flows.

2.6.2 Comparisons with Prevailing Heavy Hitter Detectors

In this section, we compare the proposed group testing approach with two prevailing sampling-based algorithms for heavy hitter detection. The first is the Sampled NetFlow algorithm introduced and implemented by Cisco [24]. Under this algorithm, one out of every r packets is sampled. If the sampled packet is from a flow that has a counter established, the counter of this flow increases by one. Otherwise, a new counter is

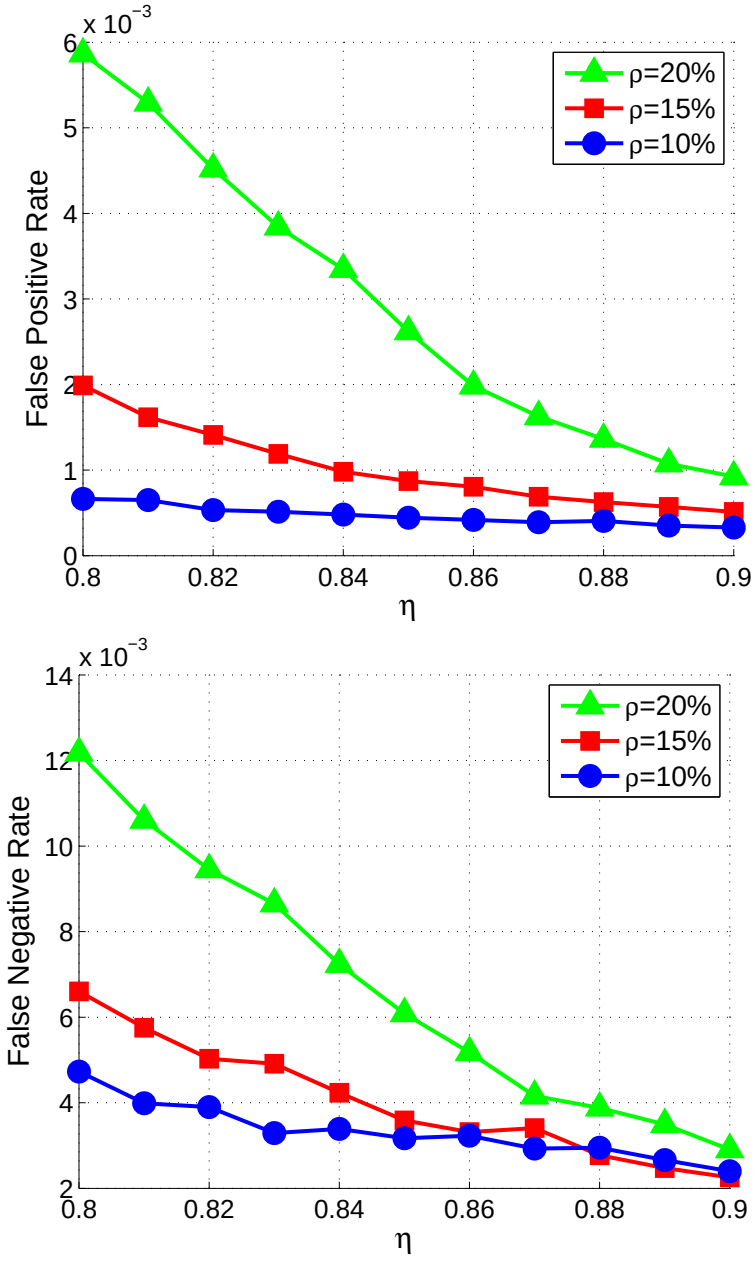


Figure 2.3: Detection accuracy of the optimal nested test plan with MLE for Poisson distributed flows ($n = 1000$, $T = 2$, $\lambda_y = 1$).

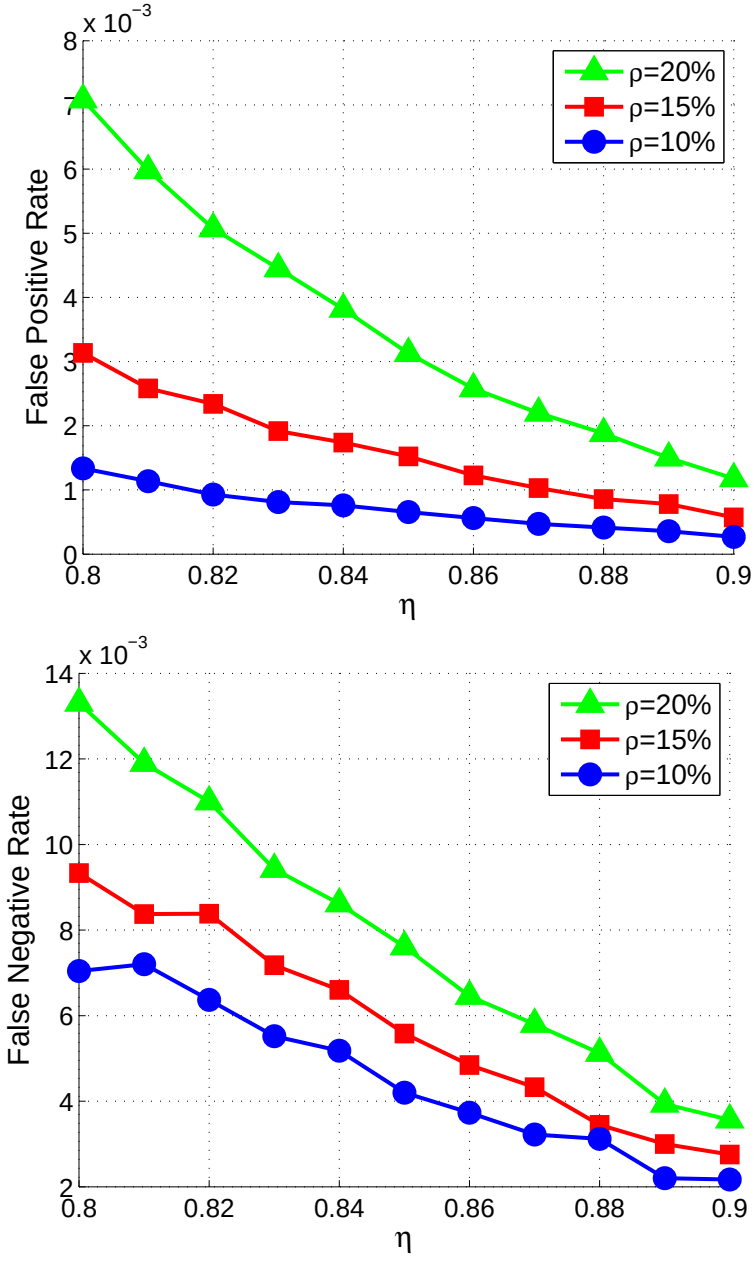


Figure 2.4: Detection accuracy of the optimal nested test plan with SME for log-normal distributed flows ($n = 1000$, $T = 5$, $\lambda_y = 1$, $\sigma_x^2 = \sigma_y^2 = 10$).

created for this flow until all available counters have been used. The sampling rate r can be chosen, often heuristically, based on the router configuration. The second algorithm is the Sample and Hold scheme introduced in [36]. Under this algorithm, the flow ID of every packet is checked. If the packet is from a flow that has a counter established, the counter of this flow increases by one. Otherwise, with probability p a new counter is created for this flow until all available counters have been used. For both algorithms, at the end of the detection window, the n_x flows with the top packet counts are declared as heavy hitters, and the rest as normal flows.

In the first example, we compare the detection accuracy as a function of the detection window of all three algorithms under a stringent counter budget. Specifically, the total number c of available counters is set to 3. For the group testing approach, c determines the maximum number of group tests that can be performed simultaneously since each group test requires counting the number of packet arrivals within an observation window of length T. The observation window T varies from 1 to 5, resulting in a detection delay (i.e., detection window) of 11 to 55 (see the x-axis of Fig. 2.5). All three algorithms are implemented over the same detection window with the same realizations of the flow processes. The parameters r and p for the two sampling-based approaches are set to their optimal values using a brute force numerical search. From Fig. 2.5 we observe that the group testing approach offers orders of magnitude improvement in detection accuracy under the same counter budget. Furthermore, the reliability of the group testing approach improves significantly when the detection window increases, while the reliability of the two sampling-based approaches remain roughly the same. This is due to the fact that a longer detection window allows a longer observation window T for each group test, thus smaller error in estimating the number of heavy hitters in each test. For the sampling-based approaches, however, detection accuracy is mainly limited by the counter budget.

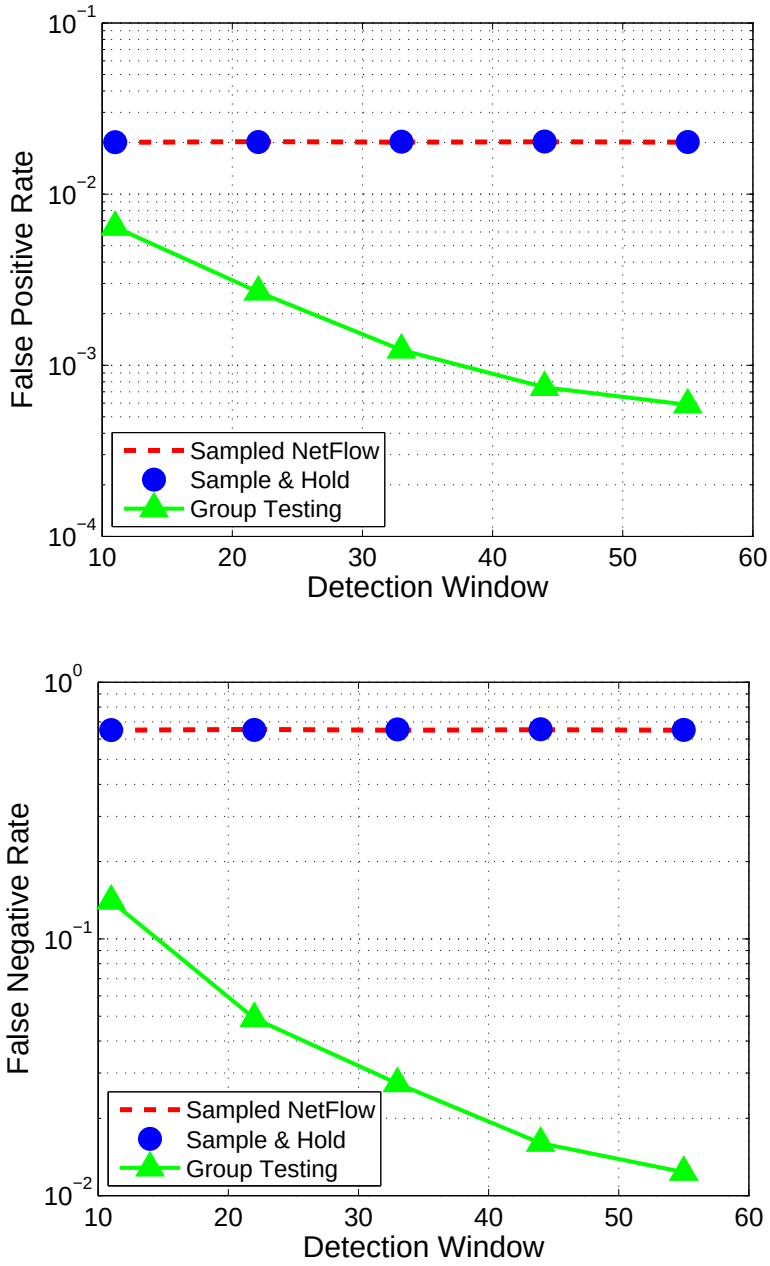


Figure 2.5: Performance comparison: detection accuracy versus detection delay ($n = 100$ Poisson flows, $n_x = 3$, $\lambda_x = 20$, $\lambda_y = 1$, $c = 3$).

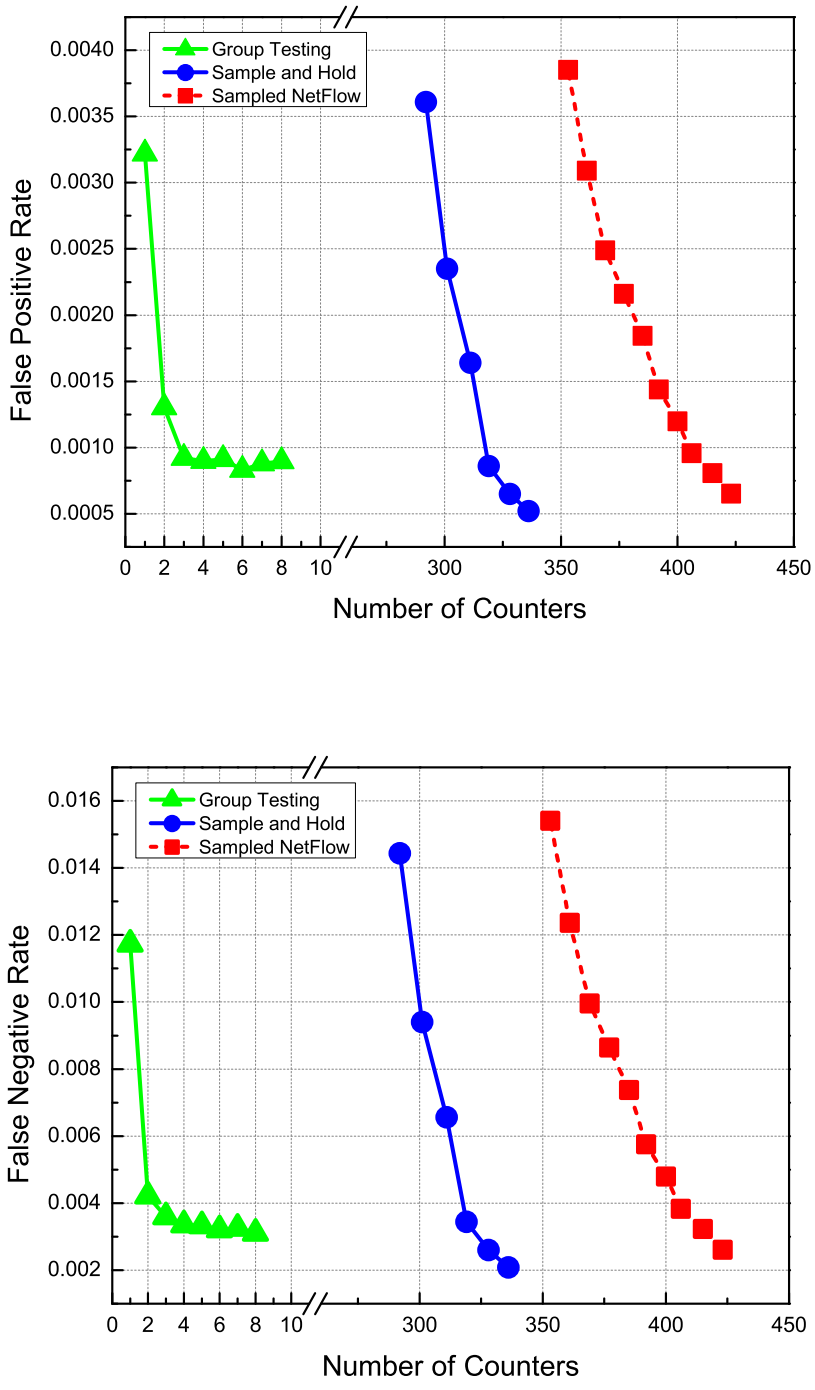


Figure 2.6: Performance comparison: detection accuracy versus counter budget ($n = 1000$ Poisson flows, $n_x = 200$, $\lambda_x = 36$, $\lambda_y = 1$, $\tau = 568$).

In the second example, we compare the counter consumption of the three algorithms by plotting the false positive and false negative rates as functions of the number of counters as shown in Fig. 2.6. The detection window τ is fixed to 568 time units for all algorithms. Again, the parameters r and p for the two sampling-based approaches are chosen optimally for each setting. For the group testing approach, the observation window T is chosen based on the counter budget so that all tests can be finished within the detection window. More specifically, with more counters, more tests can be performed simultaneously, and each test can use more observations, resulting in better detection accuracy. In particular, with a single counter, we need to set $T = 1$ in order to finish the test plan within the detection window. From Fig. 2.6 we observe that the group testing approach reduces counter consumption from hundreds to only a handful for the same level of detection accuracy.

CHAPTER 3

ACHIEVING THE OPTIMAL SCALING WITH THE SIZE OF THE SEARCH SPACE AND ACCURACY IN NOISY SCENARIOS

3.1 Background and Related Work

In this section, we study the algorithm for the anomaly detection problem of which the sample complexity achieves the optimal scaling with the size of the search space and accuracy in noisy scenarios.

3.1.1 Active Hypothesis Testing

Consider M processes among which L processes are anomalous. The decision maker aims to search for the anomalous processes by taking (aggregated) observations from a subset of processes, where the chosen subset conforms to a given tree structure. The random observations are i.i.d. over time with a general distribution that may depend on the size of the chosen subset and the number of anomalies in the subset. The objective is a sequential search strategy that adaptively determines which node on the tree to probe at each time and when to terminate the search in order to minimize the sample complexity under a constraint on the error probability.

To fully exploit the hierarchical structure of the search space, the key questions are how many samples to obtain at each level of the tree and when to zoom in or zoom out on the hierarchy. Our approach is to devise an information-directed random walk (IRW) on the hierarchy of the search space. The IRW initiates at the root of the tree and eventually arrives and terminates at the targets (i.e., the anomalous processes) with the

required reliability. Each move of the random walk is guided by the test statistic of the sum log-likelihood ratio (SLLR) collected from each child of the node currently being visited by the random walk. This local test module ensures that the global random walk is more likely to move toward a target than move away from it and that the walk terminates at a true target with the required probabilistic guarantee on detection accuracy. By constructing a sequence of last passage times of the biased random walk to shrinking subsets of the search space, we show that the sample complexity of the IRW strategy is asymptotically optimal in detection accuracy and logarithmic in M (thus order optimal as determined by the information-theoretic lower bound). The proposed search strategy is deterministic with search actions explicitly specified at each given time. It involves little online computation beyond calculating the sum log-likelihood ratio and performing simple comparisons.

3.1.2 Related Work

The anomaly detection problem considered here falls into the general class of sequential design of experiments pioneered by Chernoff in 1959 [21] in which he posed a binary (i.e., $M = 2$ for the problem at hand) active hypothesis testing problem. Compared with the classic sequential hypothesis testing pioneered by Wald [82] where the observation model under each hypothesis is fixed, active hypothesis testing has a control aspect that allows the decision maker to choose different experiments (associated with different observation models) at each time. Chernoff proposed a *randomized* strategy and showed that it is asymptotically optimal as the error probability approaches zero. Known as the Chernoff test, this randomized strategy chooses, at each time, a probability distribution governing the selection of experiments based on all past actions and observations. The probability distribution is given as a solution to a maxmin problem that can be difficult

to solve, especially when the number M of hypotheses and/or the number of experiments (which is also M for the problem at hand) is large. Furthermore, the Chernoff test does not address the scaling in M and results in a linear sample complexity in M when applied to the problem considered here. A number of variations and extensions of Chernoff's randomized test have been considered (see, for example, [9, 67, 69]). In particular, in [67], Naghshvar and Javidi developed a randomized test that achieves the optimal logarithmic order of the sample complexity in the number of hypotheses under certain implicit conditions. These conditions, however, do not hold for the problem considered here. Furthermore, similar to the Chernoff test, this randomized test is specified only implicitly as solutions to a sequence of maxmin problems that can be intractable for general observation distributions and large problem size.

The problem considered here also has intrinsic connections with channel coding with feedback, noisy group testing, and adaptive sampling with noisy response. We discuss here representative studies most pertinent to this paper and emphasize the differences in our approach from these existing studies. More detailed discussion on these related work introduced above can be found in Section 3.6.

In the channel coding with feedback, a message need to be transmitted through a channel with feedback [12, 49]. The coding problem can be reduced to an anomaly detection problem. The message that needed to be transmitted corresponds to a target among M nodes. The noisy channel can be mapped to the observation models of an active hypothesis testing problem with certain action space. i.e., Any action a with observation distribution f_a corresponds to sending a corresponding symbol through the channel and the receiving symbol at the decode end follows the distribution f_a. The IRW policy in this work provides a coding scheme with non-zero transmitting rate and asymptotically optimal error exponent for the channels including but not limited to the

discrete memoryless channel and discrete input additive noise channel with noiseless feedback.

In group testing problem, the objective is to identify defective items in a large population by performing tests on subsets of items that reveal whether the tested group contains any defective items (classic Boolean group testing) or the number of defective items in the tested group (quantitative group testing). Most work on Boolean group testing assumes error-free test outcomes. There are several recent studies on noisy group testing that assume the presence of one-sided noise [4, 76] or the symmetric case with equal size-independent false alarm and miss detection probabilities [13, 15]. The existing results on noisy group testing as well as the compressed sensing focus on non-adaptive open-loop strategies that determine all actions in one shot *a priori*. To our best knowledge, the result in this paper is the first application with adaptive test plan to noisy group testing under general noise model.

In the adaptive sampling problem [14, 22, 61, 80], the objective is estimating a step function in $[0, 1]$ or the location of an target point in $[0, 1]$ using adaptive sampling with noisy response. The main body of work on adaptive sampling is based on a Bayesian approach with binary noise of known model. A popular Bayesian strategy, the Probabilistic Bisection Algorithm, which updates the posterior distribution of the step location after each sample (based on the known model of the noisy response) and chooses the next sampling to be the median point of the posterior distribution. Several variations of the method have been extensively studied in the literature [14, 22, 61, 80]. In this work, we present a non-Bayesian approach to the adaptive sampling problem under general noise model.

The problem of detecting anomalies or outlying sequences has been studied under different formulations, assumptions, and objectives (see an excellent survey in [75] and

references therein). These studies, in general, do not address the optimal scaling in both the detection accuracy and the size of the search space. This problem is also related to the distilled sensing [48] and search with mixed observation problem [40].

3.2 Problem Formulation

We first focus on the problem of detecting a single anomalous process (referred to as the *target*) among M processes. The problem of detecting multiple targets are discussed in Section 3.5.

Let g_0 and f_0 denote, respectively, the distributions of the anomalous process and the normal processes. Aggregated observations can be obtained from a chosen subset of processes, where the subset relation is predetermined by a given tree (consider, for example, counting aggregated flows that match a given IP prefix at a programmable routers). For the ease of presentation, we focus on the case of a binary tree structure as illustrated in Fig. 3.1. Extension to a general tree structure can be found in Section 3.5.4.

Let g_l ($l = 1, \ldots, \log_2 M$) denote the distribution of the measurements that aggregate the anomalous process and $2^l - 1$ normal processes, and f_l ($l = 1, \ldots, \log_2 M$) denote the distribution of the measurements that aggregate 2^l normal processes (see Fig. 3.1). The relation between $\{g_l, f_l\}$ and $\{g_0, f_0\}$ depends on the specific application. For example, in the case of heavy hitter detection where the measurements are packet counts of an aggregated flow, g_l and f_l are given by multi-fold convolutions of f_0 and g_0. For Poisson flows, g_l and f_l are also Poisson with mean values given by the sum of the mean values of their children at the leaf level. As is the case in practically all applications, we expect that observations from each individual process are more informative than aggregated observations. More precisely, we expect $D(g_0||f_0) \geq D(g_l||f_l)$ and $D(f_0||g_0) \geq D(f_l||g_l)$

for all $l > 0$, where $D(\cdot\|\cdot)$ denotes the KL divergence between two distributions.

We aim to develop an active search strategy that sequentially determines whether to terminate the search and if not, which node on the tree to probe next. Specifically, an active search strategy $\Gamma = (\{\phi(t)\}_{t\geq 1}, \tau, \delta)$ consists of a sequence of selection rules $\{\phi(t)\}_{t\geq 1}$ governing which node to probe at each time, a stopping rule τ deciding when to terminate the search, and a declaration rule δ deciding which leaf node is the target at the time of stopping.

We adopt a Bayesian approach as in Chernoff's original work [21] and assign a cost of $c \in (0, 1)$ for each observation and a loss of 1 for a wrong declaration. Let π_m denote the *a priori* probability that process m is anomalous, which is referred to as hypothesis H_m. Let $P_e(\Gamma) = \sum_{m=1}^{M} \pi_m \alpha_m(\Gamma)$ be the probability of error under strategy Γ, where $\alpha_m(\Gamma) = \Pr_m(\delta \neq m|\Gamma)$ is the probability of declaring $\delta \neq m$ when H_m is true. Let $\mathbb{E}[\tau|\Gamma] = \sum_{m=1}^{M} \pi_m \mathbb{E}_m[\tau|\Gamma]$ be the average sample complexity of Γ. The average Bayes risk under strategy Γ is then given by

$$R(\Gamma) = P_e(\Gamma) + c\mathbb{E}[\tau|\Gamma]. \tag{3.1}$$

The objective is to find a strategy Γ that achieves the lower bound of the Bayes risk:

$$R^* = \inf_{\Gamma} R(\Gamma). \tag{3.2}$$

We are interested in test strategies that offer the optimal scaling in both c (characterizing the detection accuracy) and M. A test Γ is said to be *asymptotically optimal* in c if, for fixed M,

$$\lim_{c\to 0} \frac{R(\Gamma)}{R^*} = 1. \tag{3.3}$$

A shorthand notation $f \sim g$ will be used for $\lim_{c\to 0} f/g = 1$. A test Γ is said to be *order optimal* in c if, for fixed M,

$$\lim_{c\to 0} \frac{R(\Gamma)}{R^*} = O(1). \tag{3.4}$$

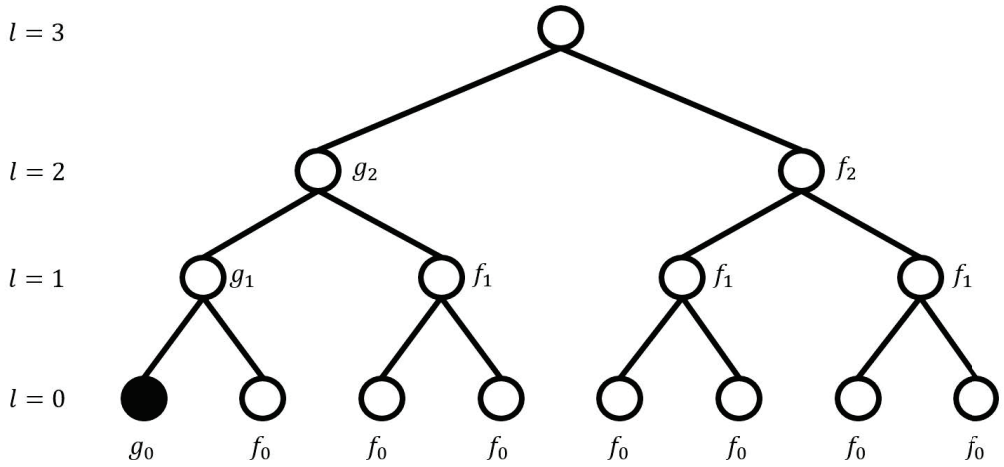


Figure 3.1: A binary tree observation model with a single target.

The asymptotic and order optimalities in M are similarly defined as the limit of M approaching infinity for all fixed c.

A dual formulation of the problem is to minimize the sample complexity subject to an error constraint ϵ, i.e.,

$$\Gamma^* = \arg\inf_{\Gamma} \mathbb{E}[\tau|\Gamma], \quad s.t. \quad P_e(\Gamma) \leq \epsilon. \tag{3.5}$$

In the Bayes risk given in (3.1), c can be viewed as the inverse of the Lagrange multiplier, thus controls the detection accuracy of the test that achieves the minimum Bayes risk. Following the same lines of argument in [59,73], one can obtain the solution of (3.5) once the solution of the Bayesian formulation is found.

3.3 Information-Directed Random Walk

In this section, we introduce the IRW policy for detection the target. The proposed policy consists of a *global random walk* on the tree interwoven with a *local test* at each

node of the tree to guide the random walk. Below we detail the two modules.

3.3.1 The Global Random Walk Module

The IRW policy induces a biased random walk that initiates at the root of the tree and eventually arrives at the target with required reliability.

Each move in the random walk is guided by the output of a local test carried on each child of the node currently being visited by the random walk. Specifically, for each run of the global random walk, assume that the policy is currently at node i on level $l > 0$ (i.e., an upper level above the leaves). If the output of the local test indicate the left (right) child contains the target, the policy zooms into the left (right) child. If the output of the local test indicates neither of the children contains the target, the policy goes back to the parent of node i. Note that we define the parent of the root node as itself. The local test module ensures that the global random walk is more likely to move toward the target than move away from it and that the random walk terminates at a true target with a sufficiently high probability.

Once arriving at a leaf node, the local test at the leaf node is performed until the policy declares the node as a target or goes back to the parent of this node.

3.3.2 The Local Test Module

To specify the local test module, suppose first that the random walk is currently at a node on a higher level $l > 0$. The objective of the local test is to distinguish three hypotheses:

$$H_0 : \text{neither of the two children contains target,}$$

$$H_1 : \text{the left child contains the target,}$$

$$H_2 : \text{the right child contains the target,}$$

correctly with probability no smaller than $\frac{1}{2}$ under each hypothesis. This problem is a miniature version ($M = 2$) of the problem described in the previous section but with the addition of hypothesis H_0.

For a node on level l, K_l samples are taken from the children of the node for deciding whether to zoom in or zoom out. The sum log-likelihood ratio (SLLR) of each child is computed independently as

$$\sum_{n=1}^{K_l} \frac{g_{l-1}(y(n))}{f_{l-1}(y(n))}, \tag{3.6}$$

where $\{y(n)\}_{n=1}^{K_l}$ represents the set of K_l samples that taken from the corresponding child.

If the SLLRs of both children are negative, The local test declares hypothesis H_0; otherwise, the local test declares H_1 (H_2) if the left (right) child has a larger SLLR. The global random walk moves based on the result of the local test. K_l is chosen to ensure that the random walk has a higher probability of moving toward than moving away from the target. Specifically, as illustrated in Fig. 3.2, at each upper level node, the random walk may go up to its parent node, go to its left child node, or go to its right child node. The probabilities for each of the three events are determined by the relative location of this node to the target and g_{l-1} and f_{l-1} (observation distributions of its children). In particular, at level l, the probability of moving closer to the target is either $p_l^{(g)}$ or $p_l^{(f)}$ depending on whether this node contains the target or not.

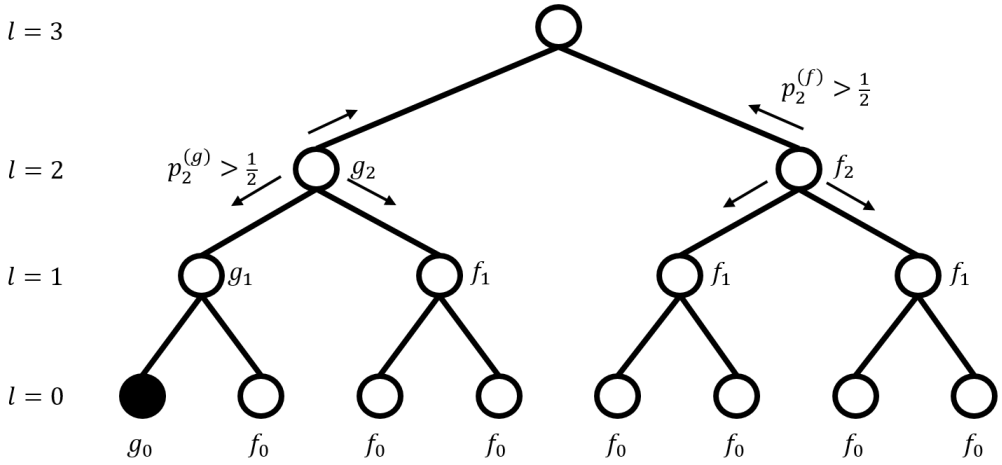


Figure 3.2: A biased random walk on the tree.

Let Y_n and Z_n denote i.i.d. random variables with distribution g_{l-1} and f_{l-1}, respectively. It is not difficult to show that $p_l^{(g)}$ and $p_l^{(f)}$ are given by

$$\begin{aligned} &\Pr\left(\sum_{n=1}^{K_l} \log \frac{g_{l-1}(Y_n)}{f_{l-1}(Y_n)} > \max\left\{\sum_{n=1}^{K_l} \log \frac{g_{l-1}(Z_n)}{f_{l-1}(Z_n)}, 0\right\}\right), \\ &\left[\Pr\left(\sum_{n=1}^{K_l} \log \frac{g_{l-1}(Z_n)}{f_{l-1}(Z_n)} < 0\right)\right]^2, \end{aligned} \tag{3.7}$$

respectively. The parameter K_l ($l = 1, 2, \ldots, \log_2 M$) is chosen as the minimum value that ensures $p_l^{(g)} > \frac{1}{2}$ and $p_l^{(f)} > \frac{1}{2}$. Note that the value of K_l can be computed offline and simple upper bounds suffice.

Once the policy arrives at a leaf node, say node m ($m = 1, \ldots, M$), samples then are drawn one by one. The SLLR of the node is computed from these samples $\{y(n)\}_{n=1}^{K}$ as

$$S_m = \sum_{n=1}^{K} \log \frac{g_0(y(n))}{f_0(y(n))}. \tag{3.8}$$

The policy continues sampling node m as long as $0 \le S_m(t) \le \log \frac{\log_2 M}{c}$. When $S_m(t)$ becomes negative, the policy goes back to the parent of node m and carries out the steps specified above for upper level nodes. The test terminates and declares the node as a target when the SLLR $S_m(t)$ of the leaf node exceeds the threshold $\log \frac{\log_2 M}{c}$.

It is a common observation that in the hypothesis testing problems the sequential tests usually have better average performance than the fixed sample size test. Inspiring by the Sequential Probability Ratio Test (SPRT), we also consider two other local tests which are the *passive* sequential local test and the *active* sequential local test. In these two local tests, instead of taking fixed number of samples from the children, the samples are taken sequentially. Similar to the SPRT, the SLLRs of the children are compared with a pair of lower and upper thresholds. Once the SLLRs satisfy the stopping criteria determined by the lower and upper thresholds, the local test stop sampling and declare whether either of the children contains the target. The probability of declaring the true hypothesis is determined by the setting of the thresholds. Details about these two sequential local tests can be found in Appendix B.1. It is shown in the numerical examples that the average performance of the IRW policy with the sequential local tests outperform the fixed-size one.

3.4 Performance Analysis of IRW Policy

We now analyze the scaling behavior of the Bayes risk of the IRW policy in terms of both M and c when there is a single target on the tree.

3.4.1 Main Structure of the Analysis

The Bayes risk of a policy consists of two parts: the sample cost and the loss associated with the detection error. For the latter, we use the union bound to bounding the detection error of the IRW policy. For the former, it consists of bounding the number of moves taken by the biased random walk and bounding the number of samples taken during

each call of the local test module.

Let D_g and D_f denote, respectively, the sojourn times at the target and a normal leaf node; they have different distributions determined by g_0 and f_0, respectively. The sample complexity of the IRW policy is analyzed by examining the trajectory of the resulting random walk. As expected, with high probability, the random walk will concentrate on a smaller and smaller portion of the tree containing the target and eventually probes the target only. Our approach is to partition the tree into $\log_2 M + 1$ half trees $\mathscr{T}_{\log_2 M}$, $\mathscr{T}_{\log_2 M-1}$, ..., $\mathscr{T}_0$ with decreasing size, and bound the time the random walk spent in each half tree. As illustrated in Fig. 3.3 for $M = 8$, $\mathscr{T}_l$ is the half tree (including the root) rooted at level l ($l = \log_2 M,\ \log_2 M - 1, \ldots,\ 1$) that does not contain the target and $\mathscr{T}_0$ consists of only the target node. The entire search process, or equivalently, each sample path of the resulting random walk, is then partitioned into $\log_2 M+1$ stages by the successively defined *last passage time* to each of the half trees in the shrinking sequence. In particular, the first stage with length $\tau_{\log_2 M}$ starts at the beginning of the search process and ends at the last passage time to the first half tree $\mathscr{T}_{\log_2 M}$ in the sequence, the second stage with length $\tau_{\log_2 M-1}$ starts at $\tau_{\log_2 M}+1$ and ends at the last passage time to $\mathscr{T}_{\log_2 M-1}$, and so on. Note that if the random walk terminates at a half tree $\mathscr{T}_l$ with $l > 0$ (i.e., a detection error occurs), then $\tau_j = 0$ for $j = l - 1, \ldots, 0$ by definition. It is easy to see that, for each sample path, we have the total time of the random walk equal to $\sum_{l=0}^{\log_2 M} \tau_l$.

Next, we consider two different scenarios regarding the quality of the aggregated observations and provide the sample complexity analysis based on the approach outlined above.

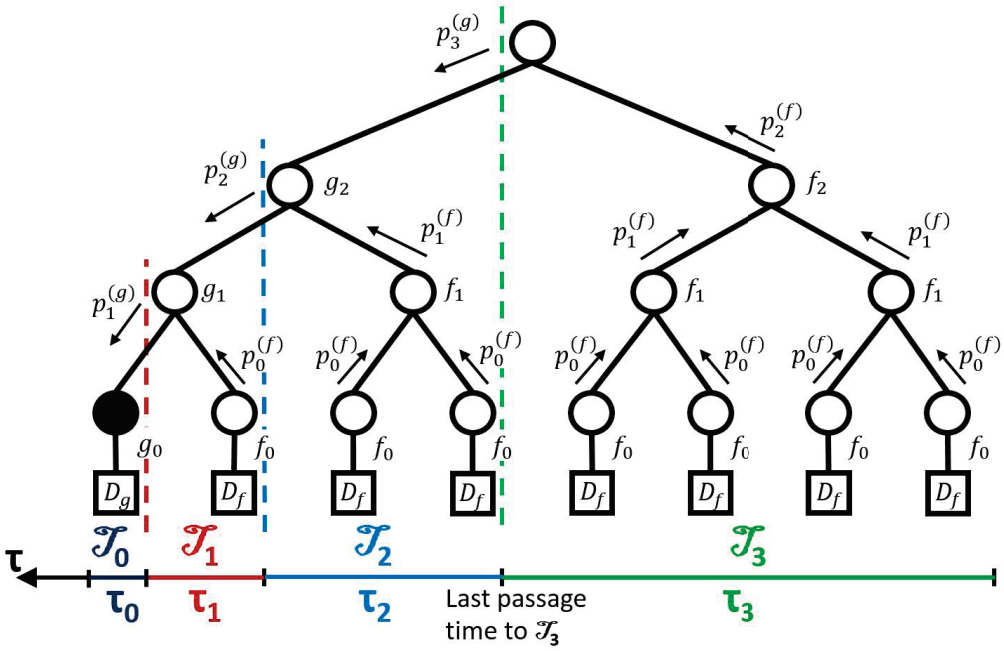


Figure 3.3: A biased random walk on the tree with sojourn times at the leaves when there is a single target.

3.4.2 Informative Observations at All Levels

We first consider the scenario where the KL divergence between aggregated observations in the presence and the absence of anomalous processes is bounded away from zero at all levels of the tree structure, i.e., there exists a constant $\delta > 0$ independent of M such that $D(g_l||f_l) > \delta$ and $D(f_l||g_l) > \delta$ for all $l = 1, 2, \ldots, \log_2 M$ and for all M. We further assume that the distributions of $\log \frac{g_0(Y_0)}{f_0(Y_0)}$ and $\log \frac{g_0(Z_0)}{f_0(Z_0)}$ are light-tailed, where Y_0 and Z_0 are random variables with the distributions g_0 and f_0, respectively. The theorem below characterizes the Bayes risk of the IRW policy.

Theorem 3. *Suppose that $D(g_l||f_l)$ and $D(f_l||g_l)$ are bounded away from zero for all l. For all M and c, we have*

$$R(\Gamma_{IRW}) \leq cB \log_2 M + \frac{c \log \frac{\log_2 M}{c}}{D(g_0||f_0)} + O(c), \tag{3.9}$$

where B is a constant independent of c and M. Furthermore, the Bayes risk of IRW is

order optimal in M for all c and asymptotically optimal in c for all M greater than a finite constant M_0.

Proof. See Appendix B.2. □

The optimality of the Bayes risk of IRW in both c and M directly carries through to the sample complexity of IRW. Specifically, from (3.9), we have the following upper bound on the sample complexity of the IRW policy

$$\mathbb{E}(\tau|\Gamma_{\text{IRW}}) \leq B \log_2 M + \frac{\log \frac{\log_2 M}{c}}{D(g_0||f_0)} + O(1). \tag{3.10}$$

We readily have

$$\mathbb{E}(\tau|\Gamma_{\text{IRW}}) \sim \frac{-\log c}{D(g_0||f_0)}, \quad \mathbb{E}(\tau|\Gamma_{\text{IRW}}) = O(\log_2 M).$$

Comparing with the lower bound developed in [25], the sample complexity of IRW is asymptotically optimal in c and order optimal in M.

3.4.3 Aggregated Observations Decaying to Pure Noise

Using Bernoulli distribution as a case study, we examine the scenario where higher level observations decay to pure noise as M grows. We establish sufficient conditions on the decaying rate of the quality of the hierarchical observations under which the proposed strategy achieves a sublinear sample complexity in M.

We assume that f_l and g_l follow Bernoulli distributions with parameters u_l and $1-u_l$, respectively. In other words, the false alarm and miss detection probabilities at level l are given by u_l. The KL divergence between g_l and f_l is $D(g_l||f_l) = D(f_l||g_l) = (1 - 2\mu_l) \log \frac{1-\mu_l}{\mu_l}$. We consider the case that μ_l increases with l and converges to 0.5 as M

approaches infinity. In this case, both $D(g_l||f_l)$ and $D(f_l||g_l)$ converge to zero, which leads to unbounded K_l. The following two theorems characterize the sample complexity of IRW when μ_l converges to 0.5 in polynomial order and exponential order, respectively.

Theorem 4. *Assume that* $\mu_l = 0.5 - (0.5 - \mu_0)(l+1)^{-\alpha}$ $(l = 0, 1, 2, \ldots, \log_2 M)$ *for some* $\alpha \in \mathbb{Z}^+$ *and* $\mu_0 < 0.5$. *The sample complexity of the IRW policy is upper bounded by:*

$$\mathbb{E}(\tau|\Gamma_{IRW}) \leq O((\log_2 M)^{2\alpha+1}) + \frac{\log \frac{\log_2 M}{c}}{D(g_0||f_0)} + O(1). \tag{3.11}$$

Proof. See Appendix B.3. □

From Theorem 4, it is not difficult to see that, for any fixed c, the IRW policy has a sample complexity that is sublinear in M:

$$\mathbb{E}(\tau|\Gamma_{\text{IRW}}) = O((\log_2 M)^{2\alpha+1}) = o(M), \text{ for } \alpha \in \mathbb{Z}^+.$$

Theorem 5. *Assume that* $\mu_l = 0.5 - (0.5 - \mu_0) \cdot \alpha^{-l}$ $(l = 0, 1, 2, \ldots, \log_2 M)$ *for some* $\alpha > 1$ *and* $\mu_0 < 0.5$. *The sample complexity of the IRW policy is upper bounded by:*

$$\mathbb{E}(\tau|\Gamma_{IRW}) \leq \tilde{B} M^{\log_2 \alpha^2} + \frac{\log \frac{\log_2 M}{c}}{D(g_0||f_0)} + O(1), \tag{3.12}$$

where $\tilde{B}$ *is a constant independent of* c *and* M.

Proof. See Appendix B.3. □

From Theorem 5, we conclude that, for any fixed c, the IRW policy has a sample complexity that is sublinear in M provided that $1 < \alpha < \sqrt{2}$. i.e.,

$$\mathbb{E}(\tau|\Gamma_{\text{IRW}}) = O(M^{\log_2 \alpha^2}) = o(M), \text{ for } 1 < \alpha < \sqrt{2}.$$

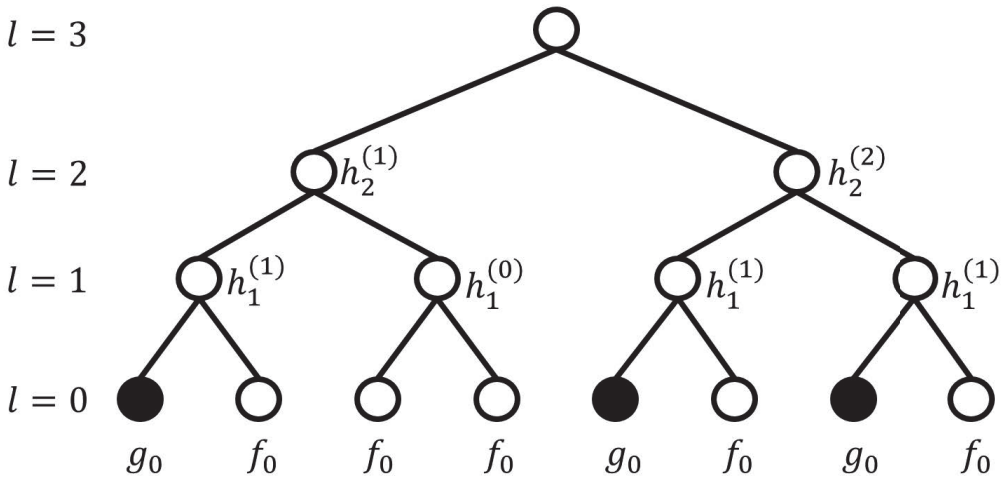


Figure 3.4: A binary tree observation model with multiple targets.

3.5 Multiple Targets and General Tree Structures

We now consider the scenarios when there are multiple targets and when the targets are on the general tree structures.

3.5.1 Formulation of Multiple Target Detection

Consider the problem of detecting L anomalous processes among M processes. A example with $M = 8$ and $L = 3$ is shown in Fig. 3.4. Let $h_l^{(d)}$ ($l = 0, 1, 2, \ldots, \log_2 M$, $d \leq \min\{L, 2^l\}$) denote the distribution of the measurements that aggregate d anomalous processes and $2^l - d$ normal processes. The objective is an active search strategy Γ for detecting the L anomalous processes. Let $H_{\mathcal{L}}$ denote the hypothesis that $\mathcal{L}$, a set of L processes, contains all the anomalous processes, and $\pi_{\mathcal{L}}$ denote the *a priori* probability of $H_{\mathcal{L}}$. The probability of error under strategy Γ is then given by $P_e(\Gamma) = \sum_{\{\mathcal{L}\}} \pi_{\mathcal{L}} \alpha_{\mathcal{L}}(\Gamma)$, where $\alpha_{\mathcal{L}}(\Gamma) = \Pr_{\mathcal{L}}(\delta \neq \mathcal{L}|\Gamma)$ is the probability of declaring $\delta \neq \mathcal{L}$ when $H_{\mathcal{L}}$ is true.

The proposed IRW policy can be easily extended to detect the multiple targets, which keeps the order optimality in M and asymptotic optimal in c. Below we discuss the details about the extended IRW policy.

3.5.2 IRW Policy for Known L

The IRW policy locates the L targets one by one[1]. Similar to the one target case, it induces a biased random walk that initiates at the root of the tree and eventually arrives at a target with required reliability (referred to as one run of the random walk). The random walk is then reset to the root until L targets have been declared.

The global random walk is guided by the local tests which indicate which of the children contains at least one undeclared targets or neither of the children contains undeclared targets. The local test module ensures that the global random walk is more likely to move toward the undeclared targets than move away from them and that the random walk arrive at a undeclared target with a sufficiently high probability.

For the local test on the upper-level nodes, when there are multiple targets on the trees, the local test becomes a *composite hypothesis testing* problem, since both the left and right children may contain more than one targets. For a node on a higher level $l > 0$, there are four hypotheses - H_0 that this node does not contain the target, H_1 (H_2) that only the left (right) child of this node contains undeclared target(s), and H_3 that both the left and right children contain undeclared targets. The objective of the local test is to correctly distinguish H_0 against H_1, H_2 and H_3 with probability no smaller than $1/2$ when H_0 is true; and declare either of the left and right children contains undeclared target(s) with probability no smaller than $1/2$ when the child truly has undeclared targets

[1]We do not assume that declared targets can be removed thus excluded from future measurements.

under either H_1, H_2 or H_3.

For the local tests of a node at level l, $K_l^{(\widehat{d})}$ are taken from the child node which already has $\widehat{d}$ declared targets, and the SLLR of this child node is computed as

$$\sum_{n=1}^{K_l^{(\widehat{d})}} \log \frac{h_{l-1}^{(\widehat{d}+1)}(y(n))}{h_{l-1}^{(\widehat{d})}(y(n))}. \tag{3.13}$$

Notice that since now $K_l^{(\widehat{d})}$ is a function of $\widehat{d}$, it may be different for the two children. Let $K_l^{(\widehat{d}_L)}$ and $K_l^{(\widehat{d}_R)}$ denote the sample sizes for the left and right children nodes, respectively.

For a given d, we assume that for any $\widehat{d} \leq d-1$,

$$D\left(h_{l-1}^{(d)} \middle\| h_{l-1}^{(\widehat{d})}\right) - D\left(h_{l-1}^{(d)} \middle\| h_{l-1}^{(\widehat{d}+1)}\right) > 0; \tag{3.14}$$

for any $\widehat{d} \geq d$,

$$D\left(h_{l-1}^{(d)} \middle\| h_{l-1}^{(\widehat{d})}\right) - D\left(h_{l-1}^{(d)} \middle\| h_{l-1}^{(\widehat{d}+1)}\right) < 0 \tag{3.15}$$

These two assumptions mean that the two distributions are more distinguishable with larger difference in the number of contained targets in them, which is usually the case in practice. If the distribution $h_l^{(d)}$ satisfies the assumptions above, the expected value of the likelihood-ratio $\log \frac{h_{l-1}^{(\widehat{d}+1)}(y(n))}{h_{l-1}^{(\widehat{d})}(y(n))}$ is positive when there exists undeclared targets on the tested child, and is negative otherwise.

Same to the one target case, after taking the specific number of samples from the two children, if the SLLRs of both children are negative, the local test declares neither of the children contains undeclared targets. Otherwise, the local test declares the child with larger SLLR contain undeclared targets. The values of $K_l^{(\widehat{d}_L)}$ and $K_l^{(\widehat{d}_R)}$ are chosen to ensure the probabilities of zooming out if neither of the two children contains undeclared targets and zooming into either of the node which contains at least one undeclared target are both greater than $1/2$. Since the number of targets (L) is finite, we can always select a K_l to make the local test be a *uniformly most powerful (UMP)* test, which guarantees

the probability of detection no matter how many undeclared targets exist on the subtree as long as it is greater than or equal to 1.

We now analyze the scaling behavior of the sample complexity and the Bayes risk of the IRW policy in terms of both M and c.

We focus on the case when observations are informative at all levels. In the other words, we consider the scenatio where the KL divergence between $h_{l-1}^{(d+k)}$ and $h_{l-1}^{(d)}$ is bounded away from zero at all levels of the tree structure, i.e., there exists a constant $\delta > 0$ independent of M such that $D(h_l^{(d+k)} || h_l^{(d)}) > \delta$ for all $l = 1, 2, \ldots, \log_2 M$, $d = 0, 1, \ldots \min\{L, 2^l\}$, and $-d \leq k \leq \min\{L, 2^l\} - d$. We still assume that the distributions of $\log \frac{g_0(Y_0)}{f_0(Y_0)}$ and $\log \frac{g_0(Z_0)}{f_0(Z_0)}$ are light-tailed, where Y_0 and Z_0 are random variables with the distributions g_0 and f_0, respectively. The following theorem characterizes the Bayes risk of the IRW policy.

Theorem 6. *Suppose that $D(h_l^{(d+k)} || h_l^{(d)})$ is bounded away from zero for all l, $d = 0, 1, \ldots \min\{L, 2^l\}$, and $-d \leq k \leq \min\{L, 2^l\} - d$. For all $M > 0$, $c < 1$, and a fixed constant $L > 0$, we have*

$$R(\Gamma_{IRW}) \leq cLB \log_2 M + \frac{cL \log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + O(c^2 \log_2 M), \tag{3.16}$$

where B is a constant independent of c and M. Furthermore, the Bayes risk of IRW is order optimal in M for all c and asymptotically optimal in c for all M sufficiently large.

Proof. We now provided the proof sketch of Theorem 6. Detailed proof can be found in Appendix B.4.

The proof is similar to the one target case which relies on the biased random walk generated by the IRW policy. As illustrated in Fig. 3.5 for the example with $M = 8$ and $L = 3$, our approach is to partition the tree into $\log_2 M + 1$ subsets. For all $l = 1, 2, \ldots, \log_2 M$, $\mathscr{T}_l$ is the union of all the nodes on level l that contains at least one

target leaf node, and their entire left or right subtree if the subtree has no target. $\mathscr{T}_0$ consists of all the target nodes. The detection process of finding any one of the targets is then partitioned into $\log_2 M + 1$ stages by the successively defined last passage time to each of these sets from upper level to lower level.

We numerate all the targets with index 1 to L from left to right. For any upper level node v on the tree, the random walk variable is defined as $D_{\min}(v) = \min_{i=1,\ldots,L}\{D_i(v)\}$, where $D_i(v)$ is the distance on the tree between current node v to the ith target. It can be shown that because of the biased random walk probability, the expected value of $D_{\min}(v)$ is always decreasing after each local test. By applying the Chernoff bound, for $l = 1, 2, \ldots, \log_2 M$, the number of samples taken on each set of the nodes is upper bounded by a constant which is independent on M. Therefore, the total sampling complexity is in logarithm order with M. The sample complexity on $\mathscr{T}_0$ equals the threshold $\log \frac{\log_2 M}{c}$ divided by $D(g_0||f_0)$, which ensures the asymptotically optimality in c. Since there are multiple targets on the tree, detection errors may have happened in previous run of IRW policy. When there are detection errors on the tree, the proof is similar but much more complex (see Appendix B.4 for details).

The sample complexity of finding each target can be bounded by $cB\log_2 M + \frac{c\log\frac{\log_2 M}{c}}{D(g_0||f_0)}$. The error probability is shown be bounded by $O(c)$. Based on this, it is not difficult to get Theorem 6 on Bayes risk. □

3.5.3 IRW Policy for Unknown L

We now extend the IRW policy for the cases when the number of abnormal processes is unknown. We consider the scenario where the KL divergence are bounded away from zero at all levels of the tree structure. Assume that the number of targets L is unknown

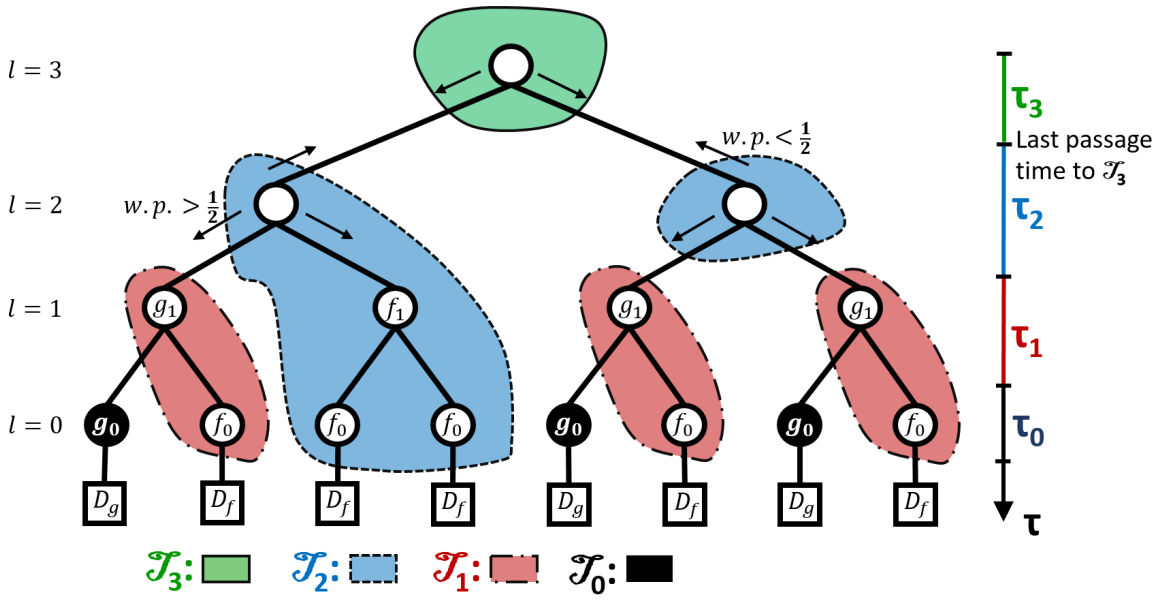


Figure 3.5: A biased random walk on the tree with sojourn times at the leaves when there are multiple targets.

but does not depend on M. The policy still locates all the targets one by one. A simple modification of the IRW policy at the root node of tree will work for the unknown number of targets.

Specifically, for each run of the IRW policy, when the random walk moves to the root node of the tree, the decision maker first carry out the local test to the two children of the root. If the local test indicates one of the children contains undeclared target, the random walk zoom into the corresponding node regularly. However, if the local test indicates neither of the children contains the undeclared targets, the policy moves into the *terminating phase* and the decision maker starts taking samples from the root node itself. The SLLR of the root node S_r is updated from zero with all its samples taken in current run of the IRW policy. The LLR of each sample $y(n)$ is computed as

$$\log \frac{h_{l_r}^{(\widehat{d}+1)}(y(n))}{h_{l_r}^{(\widehat{d})}(y(n))}, \tag{3.17}$$

where $l_r = \log_2 M$ is the level index of the root node and $\widehat{d}$ is the number of already declared targets on the tree.

The decision maker keeps sampling the root node if $\log_c \leq S_r \leq 0$. When S_r becomes less than $\log c$, the entire IRW policy terminates. When S_r becomes greater than 0, the policy moves back to the root node and carries out the regular local tests on the two children of the root until it zooms into a child or moves into the terminating phase again. For all the other tree nodes, the IRW policy works as usual as described in Section 3.3 until a targets been found and restart from the root node for next run.

It is not difficult to see that, when there are undeclared targets on the tree, the extra samples taken from the root node can be upper bounded by a constant independent of c and M. After all the targets have been found, the IRW policy takes approximately

$$\frac{-\log c}{D\left(h_{l_r}^{(L)} \middle\| h_{l_r}^{(L+1)}\right)}$$

samples before terminating. Following the similar lines of arguments in the proof of Theorem 6, the Bayes risk of the IRW policy when the number of targets is unknown can be upper bounded by

$$\begin{aligned} R(\Gamma_{\text{IRW}}) \leq & cLB \log_2 M + \frac{cL \log \frac{\log_2 M}{c}}{D(g_0 \| f_0)} \\ & + \frac{c \log \frac{1}{c}}{D\left(h_{l_r}^{(L)} \middle\| h_{l_r}^{(L+1)}\right)} + O(c^2 \log_2 M). \end{aligned} \tag{3.18}$$

3.5.4 General Tree Structures

Consider a general tree with bounded degree as shown in Fig. 3.6 as an example. We assume that each leaf of the tree follows either the distribution g_0 (target) or f_0 (non-target), although the path length from the each leaf to the root may be different. The

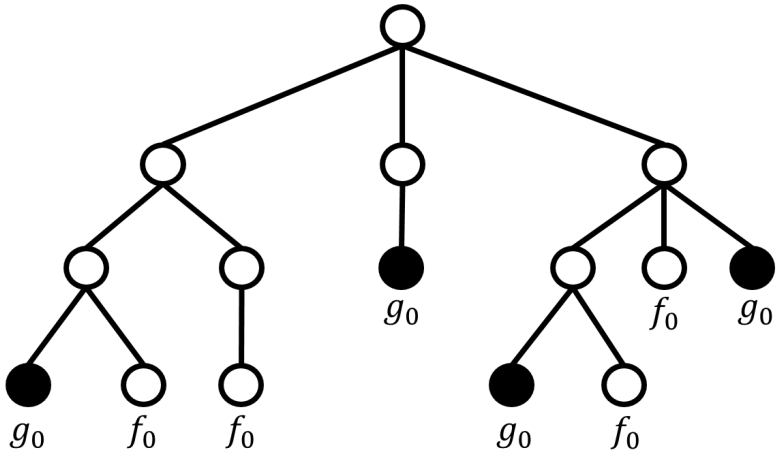


Figure 3.6: A general tree with bounded degree.

observations at any high-level nodes of the tree follows the distributions that aggregate all leaf-nodes that rooted at tested node. Let $h_{a,b}$ denote the distribution of the measurements that aggregated a anomalous processes and b normal processes.

The IRW policy for the general tree follows the similar idea. On each node of the tree, the objective of the local test is to guide the global walk zoom into the child that contains undeclared targets with probability greater than $1/2$. $K_l(\widehat{d}, w)$ samples are taken from the each child where $\widehat{d}$ is the number of declared targets and w is the number of the leaf nodes that rooted at the tested child. The SLLR of sampled child is updated as

$$\sum_{n=1}^{K_l(\widehat{d},w)} \log \frac{h_{\widehat{d}+1,w-\widehat{d}-1}(y(n))}{h_{\widehat{d},w-\widehat{d}}(y(n))}. \tag{3.19}$$

If SLLRs of all the children are negative, the local test declares no child contains undeclared target, the IRW policy goes back to the parent of current node. Otherwise, the local test declare the child which has the largest SLLR contain undeclared targets, and the IRW policy zooms into that child. $K_l(\widehat{d}, w)$'s are chosen to guarantee the probability of zooming into the children who contain undeclared targets and the probability of zooming out of the tested node if no children contains undeclared targets are both

greater than $1/2$.

Following the similar lines as in the proof of Theorem 6, we can show the Bayes risk of the IRW policy equals $O(cLHD) + O(cL\log\frac{1}{c})$, where H is height of tree, D is the maximum degree of all the tree nodes.

3.6 Discussions

We discuss below the connections between the target search problem studied in this work and several other problems that are mostly studied in different application domains. The IRW policy developed here provides an attractive solution to these problems.

3.6.1 Channel Coding with Feedback

In the channel coding with feedback, a message need to be transmitted through a channel with feedback [12, 49]. The channel state is changing over time and known at both the encoder and the decoder. As shown in Fig. 3.7 the encoder is trying transmit a message θ out of M messages to the decoder. Before any transmissions, we assume the prior of the messages is uniformly distributed to the decoder. After each transmission, the decoder updates the posterior distribution of the messages until having the required reliability to declare the true message. Due to the existing of the noiseless feedback, the encoder send symbols adaptively based on previous received symbols, of whom the objective is to transmit the message quickly and reliably.

The coding problem can be reduced to an target search problem. The message that needed to be transmitted corresponds to a target among M nodes. The noisy channel

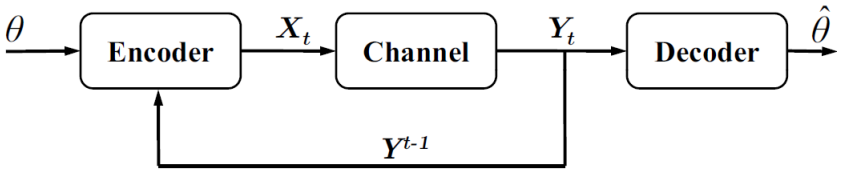


Figure 3.7: A channel coding with noiseless feedback.

can be mapped to the observation models of an active hypothesis testing problem with certain action space. i.e., Any action a with observation distribution f_a corresponds to sending a corresponding symbol through the channel and the receiving symbol at the decode end follows the distribution f_a.

We use a binary symmetric channel (BSC) with feedback as an example to show the reduction. Consider a BSC with crossover probability p. The transmission problem can be reduced to a target search on the tree problem with size-independent Bernoulli distribution observation model, where there is one target among M leaf nodes. The decision maker can take observations from each node or aggregated observations determined by three hierarchy. When the target node is contained in the tested node, the observation follows Ber$(1-p)$, i.e.,the Bernoulli distribution with probability of observing 1 equals $1-p$. If the the target node is not contained in the tested group, the observation follows Ber(p). It is not difficult to see the actions that sample the nodes contained the target is corresponding to sending 1 through the BSC, and the other actions is corresponding to sending 0 through the BSC. A policy for the AHT problem can be directly mapped to a coding scheme for the transmission problem.

The binary tree splitting generates a binary representation of the location of the target with the $\{0, 1]$ codeword of which the length equals $\log_2 M$. The test on each level of the tree corresponds to sending the next bit or correct the previous bit of the source code. Using the fixed-size local test as an example, if the left child of current node contains

the target, the sender would send K_l symbol 1's following by K_l symbol 0's through the channel. If the right child contains the target, the sender would send K_l symbol 0's following by K_l symbol 1's through the channel. If neither of the children contains the target, which means the previous bit transmitted of the source is transmitted incorrectly. In this case, the sender would send $2K_l$ symbol 0's to inform the encoder to correct the previous bit. After each local test ($2K_l$ times channel usages), a bit is sent correctly with probability greater than $\frac{1}{2}$. If a bit is sent incorrectly, it can also be revisited and corrected later with probability greater than $\frac{1}{2}$. When the policy arrives at a leaf node, since the full codeword has been transmitted, the sender will keep sending symbol 1 if the codeword is correct at the received end until the log-likelihood ratio is large enough. If the codework is transmitted incorrect, the sender will keep sending 0 until the receiver correct the belief. The step of sending the confirmation bits corresponds to the local test on a leaf in the IRW policy.

Following the similar approach, the coding problems over any stationary Discrete Memoryless Channel (DMC) and any stationary discrete-input additive noise channel with noiseless feedback can be reduced to an target search on a tree considered in this work. A stationary DMC with a finite number of input symbols $I_1, I_2, \ldots, I_J$ and output symbols $O_1, O_2, \ldots, O_K$ is defined by the transition probability matrix $\mathbb{P} = \{p_{j,k}\}$, $j = 1, \ldots, J$, $k = 1, \ldots, K$. When it is reduced to the target search on the tree, one of the leaves is the target. The distribution of the target g_0 is set as the probability mass vector $\{p_{j_g^*,k}\}$, $k = 1, \ldots, K$ and the distribution of non-target node f_0 is set as the probability mass vector $\{p_{j_f^*,k}\}$, $k = 1, \ldots, K$, where j_g^* and j_f^* are defined as

$$\left(j_g^*, j_f^*\right) = \arg\max_{(j_g, j_f)} \sum_{k=1}^{K} p_{j_g,k} \log \frac{p_{j_g,k}}{p_{j_f,k}}, \forall j_g, j_f = 1, \ldots, J. \tag{3.20}$$

The observation distributions g_l, f_l from the nodes on level $l \geq 1$ of the tree can be arbitrarily mapped from the probability mass vectors $\{p_{j_g,k}\}$ and $\{p_{j_f,k}\}$, with $k =$

$1, \ldots, K$, $j_g \neq j_f$.

Similarly, a stationary discrete-input additive noise channel can also be reduced to the target search problem on a binary tree. After such reduction, the proposed IRW policy provided a coding scheme with non-zero transmit-rate that achieves the optimal error exponent. In order to achieve the optimal transmit rate (capacity of the channel), a careful mapping of the observation on the high-level nodes of the tree need to be designed, which is an interesting future direction.

Another related problem is the problem of reliably transmitting a real-valued random vector through a digital noisy channel [26]. For simplicity, consider the problem of transmitting a real-valued point x on the interval $[0, 1]$ through a digital noisy channel. The transmission scheme consists of an encoder and a sequential decoder. The encoder is a family of maps $E_t : \theta \to \mathcal{X}$ specifying the symbol transmitted through the channel at time t. The decoder is a family of maps $D_t : \mathcal{Y}_t \to \theta$, describing the estimate $\hat{\theta}_t = D_t((y_s)_{s=1}^t)$ of θ from the sequence $(y_s)_{s=1}^t$ that has been received through the channel until time t. The objective is to find a transmission scheme to make the root mean squared error converge to zero with degree α and rate β. i.e.,

$$\Delta_t := (\mathbb{E}[||\theta - \hat{\theta}_t||^2])^{1/2} \leq p(t)2^{-\beta t^\alpha}, \tag{3.21}$$

for some constants $\beta > 0$ and $0 < \alpha \leq 1$. Similarly, the proposed IRW policy can be applies to transmit the point on the interval $[0, 1]$. The binary tree structure is a Tree-Structured Vector Quantization of the real-point [41]. It has been shown that if $Q : \mathcal{X} \to \mathcal{X}$ is a quantizer assuming m values then

$$(\mathbb{E}[||\theta - Q(\theta)||^2])^{1/2} \geq Cm^{-1}. \tag{3.22}$$

This shows that $\Delta_t \geq C2^{-t}$. We apply the IRW policy to the case with binary symmetric channel with crossover probability 0.3. Fig.3.8 shows the order optimality of the root mean squared error when $M = 2^{25}$.

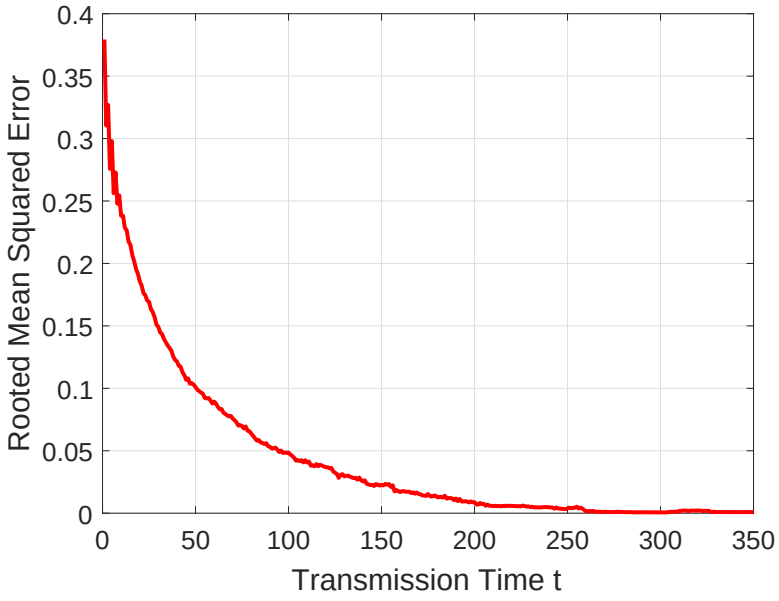


Figure 3.8: Root mean squared error of IRW for point transmission through a B-SC.

3.6.2 Noisy Group Testing and Compressed Sensing

In the group testing problem, the objective is to identify defective items in a large population by performing tests on subsets of items that reveal whether the tested group contains any defective items (classic Boolean group testing) or the number of defective items in the tested group (quantitative group testing). Most work on Boolean group testing assumes error-free test outcomes. There are several recent studies on noisy group testing that assume the presence of one-sided noise [4, 76] or the symmetric case with equal size-independent false alarm and miss detection probabilities [13, 15]. In some extended group testing models such as the noisy quantitative group testing [83] and threshold group testing [27], the issue of sample complexity in terms of detection accuracy is absent in the basic formulation. Similar to the group testing problem, the objective of the compressed sensing problem [4] is to recover a sparse signal with aggregated observations. The existing results on noisy group testing as well as the compressed sensing

focus on non-adaptive open-loop strategies that determine all actions in one shot *a priori*. The disadvantages of non-adaptive test plans lie in the computational complexity of the coding/decoding processes and high storage requirement.

The above various formulations of group testing and the compressed sensing problems can be reduced to the active search problem with specific observation distributions (e.g., Bernoulli distribution for noisy Boolean group testing, sum-observation model for the quantitative and threshold group testing). The proposed IRW policy provides a *sequential* and *adaptive* solutions to solve the group testing and compressed sensing with little offline or online computation. The inherent tree structure of the test plan also leads to low memory requirement. It is thus particularly attractive for online applications such as real-time heavy hitter detection where n and d are not prefixed and computational, memory, and counter resources are stringent. More importantly, the policy works for general noisy observation models. Although adaptive group testing strategies do not necessarily conform to a predetermined tree structure, the proposed IRW policy offers asymptotic optimality in both population size and reliability constraint.

3.6.3 Adaptive Sampling with Noisy Response

In the adaptive sampling problem [14, 22, 61, 80], the objective is estimating a step function in $[0, 1]$ or the location of an target point in $[0, 1]$ using adaptive sampling with noisy response. We limit the inpus space to be one-dimensional in order to demonstrate the main idea. The main body of work on the adaptive sampling is based on a Bayesian approach with binary noise of a known model. A popular Bayesian strategy, the Probabilistic Bisection Algorithm, which updates the posterior distribution of the step location after each sample (based on the known model of the noisy response) and chooses the

next sampling to be the median point of the posterior distribution. Several variations of the method have been extensively studied in the literature [14, 22, 61, 80]. However, the size of the search space is extremely large when $M \to \infty$, especially when there are multiple targets ($L > 1$) the size of possible hypotheses space grows exponentially with L. After taking each sample, the update of the posterior beliefs as well as the sorting of the posterior distribution is extremely costly.

Partitioning the $[0, 1]$ interval into small intervals and sampling the group of intervals, we can map the adaptive sampling problem to the target search problem where the target is the small interval containing the location of the target.

For problem of estimating a step function in $[0, 1]$, the hypothesis class, denoted by $\mathcal{H}$, is the set of all step functions on $[0, 1]$ interval

$$\mathcal{H} = \{h_z : [0, 1] \to \mathbb{R}, h_z(x) = \mathbb{I}_{\{(z,1]\}}(x), z \in (0, 1)\}, \tag{3.23}$$

where

$$\mathbb{I}_{\{(z,1]\}}(x) = \begin{cases} 0, & \text{when } x \leq z, \\ 1, & \text{when } x > z. \end{cases}$$

Each hypothesis h_z assigns a binary label to each element of the input space $[0, 1]$. There is a true hypothesis h_z^* that determines the ground truth labels for the input space. The learner is allowed to make sequential observations by adaptively sampling h_z^*. The observations are however noisy. The goal is to design a sequential sampling strategy aiming at minimizing the sample complexity required to obtain a confidence interval of length δ for z^* with required reliability. Specifically, the learner chooses the sampling point x at each time t and receives a noisy sample of the true hypothesis.

Without loss of generality, we assume $\delta = 1/M$, where M is a power of two. Let each leaf node on a binary tree represent an interval $\left[\frac{i}{M}, \frac{i+1}{M}\right]$ for $i = 0, 1, \ldots, M-1$. The

interval corresponding to each upper-level node on the tree is the union of the intervals corresponding to its children. Examining larger intervals (consisting of several smaller intervals) induces a hierarchical structure of the noisy responses. The proposed IRW policy can be applied to search the interval contains the target. In the local test module, the test on a child node is corresponding to the test that takes samples from the left and right boundaries of the interval, of which the objective is to determine whether the observation from the left boundary is 0 and from the right boundary is 1, i.e., the target point is in the current interval. The probability of zooming into the intervals is guaranteed to be greater than $\frac{1}{2}$. The proposed IRW strategy is deterministic with search actions explicitly specified at each given time provides an efficient way to solve the adaptive sampling problem. It involves little online computation beyond calculating the sum log-likelihood ratio and performing simple comparisons.

3.7 Simulation Examples

We now provide the numerical examples of the IRW policy as well as the comparison with the Chernoff test and the DGF test developed in [25].

In this example, we consider detecting L heavy hitters among Poisson flows and the measurements are exponentially-distributed packet inter-arrival times. For the leaf-node, g_0 and f_0 are exponential distributions with parameters λ_g and λ_f, respectively. The aggregated flows follow the corresponding exponential distributions with the parameters equal to the sum of the parameters of their children at the leaf level.

Under the same action space given by all nodes on the tree, the resulting Chernoff test, however, probes only the leaf nodes. Specifically, at each time t, all the leaf nodes

are sorted based on their SLLRs. If

$$D(g_0||f_0)/L \geq D(f_0||g_0)/(M-L), \tag{3.24}$$

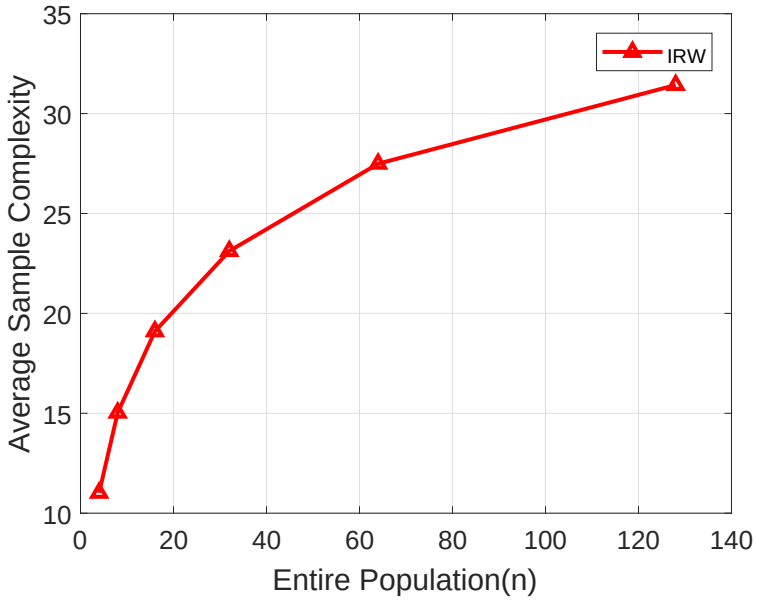
the Chernoff test randomly and uniformly selects one node from the ones with the largest SLLR to the Lth largest SLLR; if

$$D(g_0||f_0)/L < D(f_0||g_0)/(M-L), \tag{3.25}$$

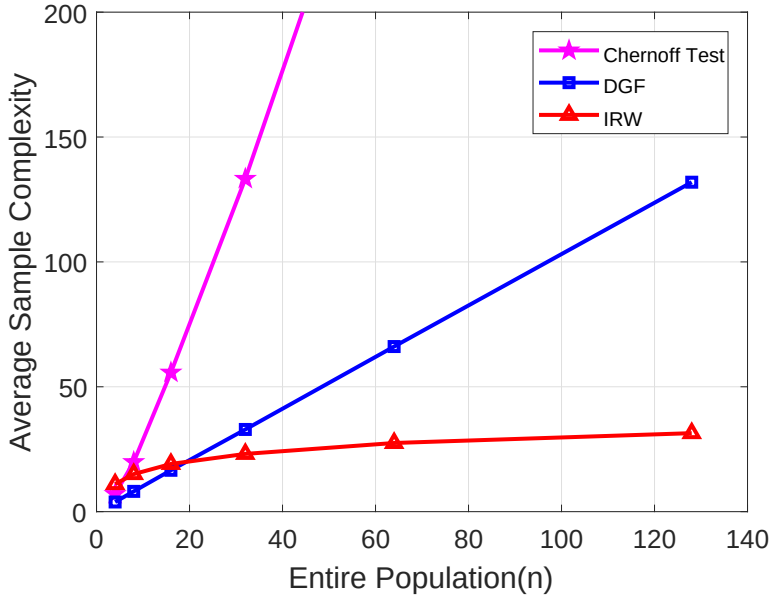
the Chernoff test randomly and uniformly selects one node from the ones with the $(L+1)$th largest SLLR to the smallest SLLR. Under condition (3.24), the DGF test probes the node with the Lth largest SLLR. Under condition (3.25), the DGF test probes the node with the $(L+1)$th largest SLLR. Both the Chernoff test and the DGF test update the SLLR of leaf nodes with each corresponding sample and terminate when the SLLR difference between the Lth largest and the $(L+1)$th largest ones exceeds the threshold $-\log c$; then declare the node with the largest SLLR as the target.

Fig. 3.9 shows a simulation example comparing the sample complexity for detection a single target on the tree as a function of M of the Chernoff test, the DGF policy, and the IRW policy with the fixed sample size local test. The sample complexity of the Chernoff test and the DGF test increase linearly with M, while the sample complexity of the IRW policy increases in a logarithmic-order with M. The number of sampling in each local tests is relative large comparing to M when M is small, therefore IRW does not outperform DGF at the beginning. However, the advantage of the IRW policy in terms of the sample complexity is significant as M increases.

As introduced in Section 3.3.2, we introduced another two sequential versions of the local tests. The sequential local tests make the performance of the IRW policy even better. We now compare the sampling complexity of IRW policy with fixed-size local test, passive and active sequential local tests in numerical examples. We assume there



(a) IRW



(b) Comparison

Figure 3.9: Performance comparison of three test plans ($L = 1$, $\lambda_g = 10, \lambda_f = 0.01, K_l = 3, c = 10^{-13}$, $n = 8, 16, \ldots, 1024$.).

is one target on the tree and the observation on each level of the tree follows Bernoulli distribution with probabilities 0.6 or 0.4 to observing 1 when the node contains the target

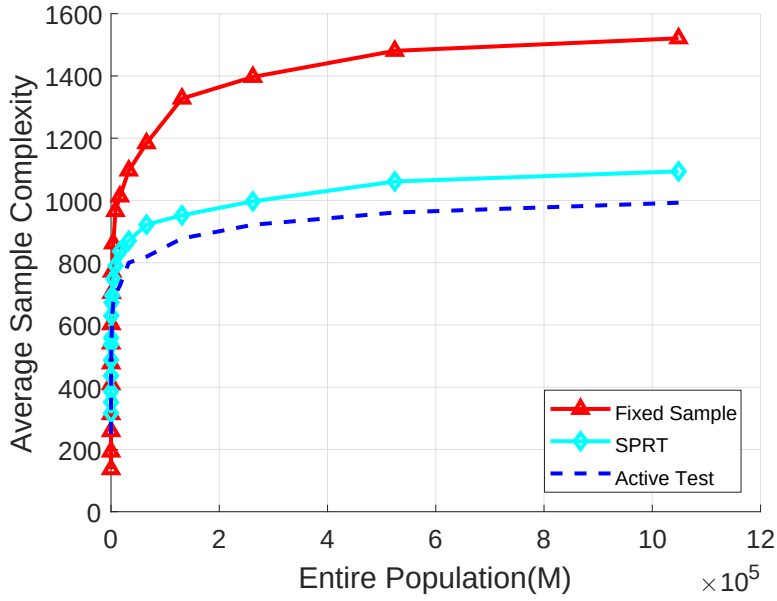
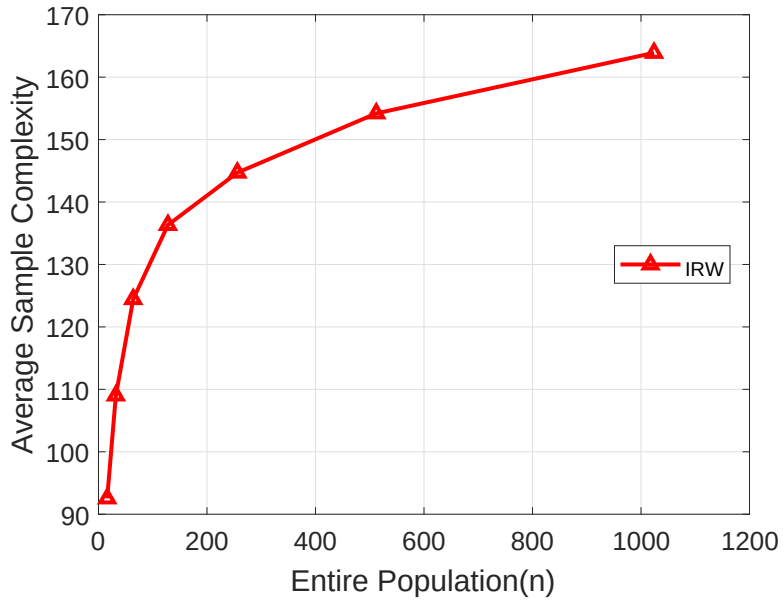


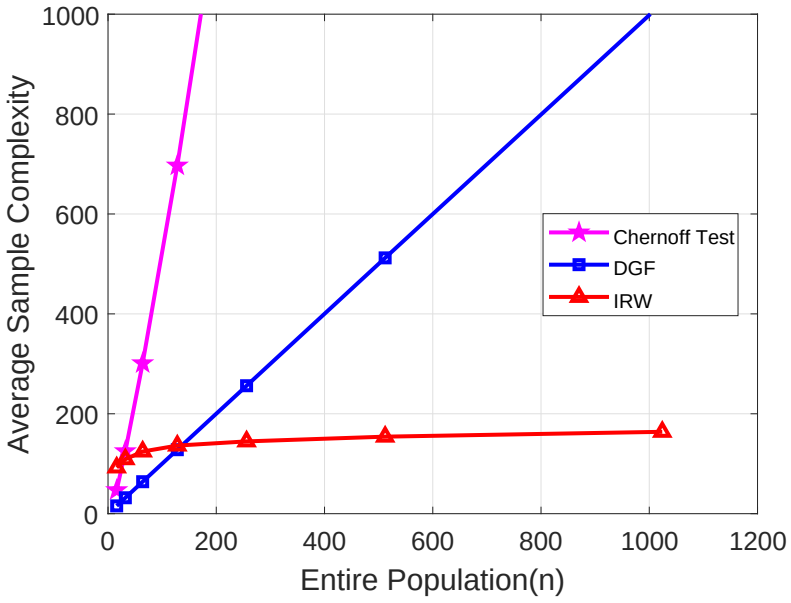
Figure 3.10: Performance comparison of IRW with three different local tests. ($K_l = 7$, $\gamma_1 = 1.0986$, $\gamma_0 = -1.0986$, $\nu_1 = 0.9445$, $\nu_0 = -0.9445$, and 1000 Monte Carlo runs.)

or not, respectively. For all the three local tests, we set the probability of approaching the target to be 0.5625. The value of K_l for the fixed-size local test, γ_0 and γ_1 for the passive sequential test, and ν_0 and ν_1 are set accordingly. The comparison is shown in Fig. 3.10 where we can find that the two sequential local tests have much better performance than the fixed size local test and the active one has the best performance among all the three.

Besides the saving in the sample complexity, the sequential local tests are also easier to implement. Instead of designing K_l for each level, one pair of thresholds for the SLLR that guarantee the biased global random walk can be use on all the higher level nodes. Fig. 3.11 shows a simulation example comparing the sample complexity for detection five targets on the tree of the Chernoff test, the DGF policy, and the IRW policy with the active sequential local test. Similar to the single target case, the IRW policy outperforms the other two.



(a) IRW



(b) Comparison

Figure 3.11: Performance comparison of three test plans ($L = 5$, $\lambda_g = 10, \lambda_f = 0.001, c = 5 \times 10^{-5}$, $n = 8, 16, \ldots, 1024$.).

CHAPTER 4

CONCLUSION

This dissertation focuses on the sequential design of experiments for anomaly detection. Specifically, the problem of detecting a few anomalous processes among a large number of processes is considered.

In the first part of the dissertation, we develop an algorithm for the anomaly detection of which the sample complexity is in optimal scaling with the size of the search space. We consider the case where the observations from all the processes are noiseless. We studied the quantitative group testing problem within the combinatorial group testing framework. The optimal nested test plan was established in closed form. Its application in heavy hitter detection was studied and its performance compared with prevailing sampling-based approaches.

In the second part of the dissertation, we develop an algorithm for the anomaly detection problem of which the sample complexity is in optimal scaling with the size of the search space as well as the accuracy requirements. We consider the case where the observations from the processes are noisy. The sample complexity of the proposed active search strategy with the global random walk and local test modules is shown to be order optimal with the size of the search space under certain conditions on the observation model and asymptotic optimality in terms of the reliability constraint. In this work, the stochastic models of all processes are assumed known. An extension of the IRW policy to anomaly detection under unknown models can be found in [79].

APPENDIX A

PROOF FOR LEMMAS AND THEOREMS IN CHAPTER 2

A.1 Proof of Lemma 1

Consider first the initial value of each sequence with $n = 2d$. This corresponds to $l = 0$, $k = d$. Setting $t = 2$, $i = 0$ in (2.9), we have

$$N(2d, d) \leq 2d - 1.$$

Together with (2.8), we arrive at

$$N(2d, d) = (l + 1)d + k - 1 = 2d - 1.$$

When $n > 2d$, based on the definition of l and k, we write n as

$$n = (d + k - 1)2^l + (n - (d + k - 1)2^l). \tag{A.1}$$

Let $x = n - (d + k - 1)2^l$. It is easy to see that $l \geq 1$, $1 \leq k \leq d$, and $1 \leq x \leq 2^l$. Based on (A.1), (2.5) is equivalent to

$$N((d + k - 1)2^l + x, d) = (l + 1)d + k - 1. \tag{A.2}$$

Setting $t = l + 1$, $i = k - 1$ in (2.10), we have

$$N((d + k - 1)2^l + 1, d) \geq (l + 1)d + k - 1. \tag{A.3}$$

Setting $t = l + 1$, $i = k$ for $1 \leq k \leq d - 1$ and $t = l + 2$, $i = 0$ for $k = d$ in (2.9), we have,

$$N((d + k - 1)2^l + 2^l, d) \leq (l + 1)d + k - 1. \tag{A.4}$$

From (A.3) and (A.4) we have, for all $x = 1, 2, \ldots, 2^{t-1}$,

$$\begin{aligned}(l+1)d + k - 1 &\leq N((d+k-1)2^l + 1, d) \\ &\leq N((d+k-1)2^l + x, d) \\ &\leq N((d+k-1)2^l + 2^l, d) \\ &\leq (l+1)d + k - 1,\end{aligned}$$

which leads to (A.2). Here we have used the monotonicity property of $N(n, d)$, i.e., $N(n, d) \leq N(n+1, d), \forall n \geq 2d$.

A.2 Proof of Properties of $N(n, d)$

In this appendix, we provide the proof of the three properties [P1]-[P3] introduced in Section 2.4.

A.2.1 Proof of [P1]

The proof is based on induction in n using the recursive formulas in (2.3) and (2.4).

Let $d_1^*(m;\, n, d)$ denote the maximizer that achieves $\phi(m;\, n, d)$ as defined in (2.16), i.e.,

$$\phi(m;\, n, d) = N(m, d_1^*(m;\, n, d)) + N(n-m, d - d_1^*(m;\, n, d)). \tag{A.5}$$

Note that since d and m are restricted to no greater than $\frac{n}{2}$, we have $0 \leq d_1^*(m;\, n_0, d) \leq \min\{m, d\}$.

The initial condition of the induction is easy to check: $N(2, 1) = 1 > N(2, 0) = 0$. Now assume that there exists an $n_0 > 2$ such that for every $n < n_0$, $\{N(n, d)\}_{d=0}^{\lfloor n/2 \rfloor}$ is a

strictly increasing sequence in d. Based on this induction assumption, we prove next that $\{N(n_0, d)\}_{d=0}^{\lfloor n_0/2 \rfloor}$ is strictly increasing in d.

It is straightforward that $N(n_0, 0) < N(n_0, 1)$. When $d > 2$, we prove the statement by considering separately the cases when n_0 is odd and when n_0 is even.

Case 1: n_0 is odd.

The basic idea of the proof is to show that for all $m = 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$,

$$\phi(m;\, n_0, d-1) < \phi(m;\, n_0, d). \tag{A.6}$$

Then from (2.3), we arrive at [P1].

Next, we show (A.6) by considering the following two cases in terms of the value of $d_1^*(m;\, n_0, d-1)$:

$$0 \le d_1^*(m;\, n_0, d-1) < \min\left\{ \left\lfloor \frac{m}{2} \right\rfloor, d \right\}, \tag{A.7}$$

$$d - 1 - \left\lfloor \frac{n_0 - m}{2} \right\rfloor < d_1^*(m;\, n_0, d-1) \le \min\{m,\, d-1\}. \tag{A.8}$$

It is easy to see that (A.7) and (A.8) cover all possible values of $d_1^*(m;\, n_0, d-1)$ since the upper limit in (A.7) is greater than the lower limit in (A.8) given that $m \le \left\lfloor \frac{n_0}{2} \right\rfloor$ and $d \le \left\lfloor \frac{n_0}{2} \right\rfloor$.

When (A.7) is true, we have

$$\begin{aligned}
\phi(m;\, n_0, d) &\overset{(a)}{\ge} N(m, d_1^*(m;\, n_0, d-1) + 1) + N(n_0 - m, d - d_1^*(m; n_0, d-1) - 1) \\
&\overset{(b)}{>} N(m, d_1^*(m;\, n_0, d-1)) + N(n_0 - m, d - d_1^*(m; n_0, d-1) - 1) \\
&\overset{(c)}{=} \phi(m;\, n_0, d-1),
\end{aligned}$$

where (a) holds since $d_1^*(m;\, n_0, d-1) + 1$ is in the range $\{0, \ldots, \min\{m, d\}\}$ of the maximizer for $\phi(m;\, n_0, d)$; (b) follows from the induction hypothesis and the fact that $d_1^*(m;\, n_0, d-1) < \left\lfloor \frac{m}{2} \right\rfloor$ given in (A.7), and (c) follows from (A.5). We thus arrive at (A.6).

When (A.8) is true, by noticing that $d_1^*(m;\, n_0, d-1)$ is within the range $\{0, \ldots, \min\{m, d\}\}$ of the maximizer for $\phi(m;\, n_0, d)$, we have

$$\begin{aligned}\phi(m;\, n_0, d) &\geq N(m, d_1^*(m;\, n_0, d-1)) + N(n_0 - m, d - d_1^*(m; n_0, d-1)) \\ &> N(m, d_1^*(m;\, n_0, d-1)) + N(n_0 - m, d - d_1^*(m; n_0, d-1) - 1) \\ &= \phi(m;\, n_0, d-1).\end{aligned}$$

This concludes the proof for *Case 1*.

Case 2: n_0 is even.

For $d < \frac{n_0}{2}$, the proof follows the same line of argument as in *Case 1*. Now consider $d = \frac{n_0}{2}$. We need to prove $N(n_0, \frac{n_0}{2} - 1) < N(n_0, \frac{n_0}{2})$. Base on Lemma 1, we have $N(n_0, \frac{n_0}{2}) = n_0 - 1$. Then it is equivalent to prove $N(n_0, \frac{n_0}{2} - 1) < n_0 - 1$.

When m is even, $d_1^*(m;\, n_0, \frac{n_0}{2} - 1)$ is covered by (A.7) and (A.8). The same line of arguments as in *Case 1* leads to

$$\phi(m;\, n_0, \frac{n_0}{2} - 1) < \phi(m;\, n_0, \frac{n_0}{2}). \tag{A.9}$$

Based on the unimodal property of $\{N(n, d)\}_{d=0}^{n}$ given in [P3], we further have

$$d_1^*(m;\, n_0, \frac{n_0}{2}) = \frac{m}{2}, \tag{A.10}$$

i.e.,

$$\phi(m;\, n_0, \frac{n_0}{2}) = N(m, \frac{m}{2}) + N(n_0 - m, \frac{n_0 - m}{2}) \tag{A.11}$$

$$= n_0 - 2, \tag{A.12}$$

where (A.12) is based on Lemma 1. Therefore, we have, for all even m,

$$\phi(m;\, n_0, \frac{n_0}{2} - 1) < n_0 - 2. \tag{A.13}$$

When m is odd, based on [P3] and Lemma 1, we have

$$\phi(m;\, n_0, \frac{n_0}{2} - 1) = N(m, \frac{m-1}{2}) + N(n_0 - m, \frac{n_0 - m - 1}{2}) = n_0 - 2. \tag{A.14}$$

With (A.13) and (A.14), we have

$$N(n_0, \frac{n_0}{2} - 1) = 1 + \min_m \phi(m;\, n_0, \frac{n_0}{2} - 1) < n_0 - 1,$$

i.e.,

$$N(n_0, \frac{n_0}{2} - 1) < N(n_0, \frac{n_0}{2}). \tag{A.15}$$

A.2.2 Proof of [P2]

We first establish the following lemma.

Lemma 2. *Let $f(x)$ be a real-valued function defined on a finite set of consecutive integers, i.e., $x \in \{u, u+1, \ldots, v\}$ for some u and v. Suppose that $f(x)$ is monotonically increasing and concave. For every positive integer s, let $\{c_k\}_{k=0}^{s}$ be an arbitrary increasing and concave sequence. Define, for $x = u, u+1, \ldots, v+s$,*

$$F(x) := \max\{f(x) + c_0, f(x-1) + c_1, \ldots, f(x-\tau) + c_\tau\},$$

where $\tau = \min\{x - u, s\}$. Then $F(x)$ is increasing and concave.

This lemma is rather intuitive given that $F(x)$ is the maximum of shifted versions of $f(x)$ which is increasing and concave. An numerical example with $s = 3$ is given in Fig. A.1.

Next we provide a detailed proof of this lemma. Our objective is to show

$$0 \le F(x+1) - F(x) \le F(x) - F(x-1). \tag{A.16}$$

Based on the definition of $F(x)$, we have

$$F(x-1) = \max_{k=0,1,\ldots,\min\{x-u-1,s\}} \{f(x-k-1)+c_k\},$$

$$F(x) = \max_{k=0,1,\ldots,\min\{x-u,s\}} \{f(x-k)+c_k\},$$

$$F(x+1) = \max_{k=0,1,\ldots,\min\{x-u+1,s\}} \{f(x-k+1)+c_k\}.$$

Define k^* as

$$k^* = \arg \max_{k=0,1,\ldots,\min\{x-u,s\}} \{f(x-k)+c_k\},$$

i.e.,

$$F(x) = f(x-k^*)+c_{k^*}.$$

Based on this definition, we have

$$f(x-k^*) - f(x-(k^*+1)) \geq c_{k^*+1} - c_{k^*}, \tag{A.17}$$

$$f(x-(k^*-1)) - f(x-k^*) \leq c_{k^*} - c_{k^*-1}. \tag{A.18}$$

Based on (A.17) and (A.18), due to the monotonicity and the concavity of $f(x)$ and $\{c_k\}_{k=1}^n$, we can easily prove

$$\begin{aligned} f(x-(k^*+1)) - & f(x-(k^*+2)) \\ & \geq f(x-k^*) - f(x-(k^*+1)) \\ & \geq c_{k^*+1} - c_{k^*} \\ & \geq c_{k^*+2} - c_{k^*+1} \end{aligned} \tag{A.19}$$

$$\begin{aligned} f(x-(k^*-2)) - & f(x-(k^*-1)) \\ & \leq f(x-(k^*-1)) - f(x-k^*) \\ & \leq c_{k^*} - c_{k^*-1} \\ & \leq c_{k^*-1} - c_{k^*-2} \end{aligned} \tag{A.20}$$

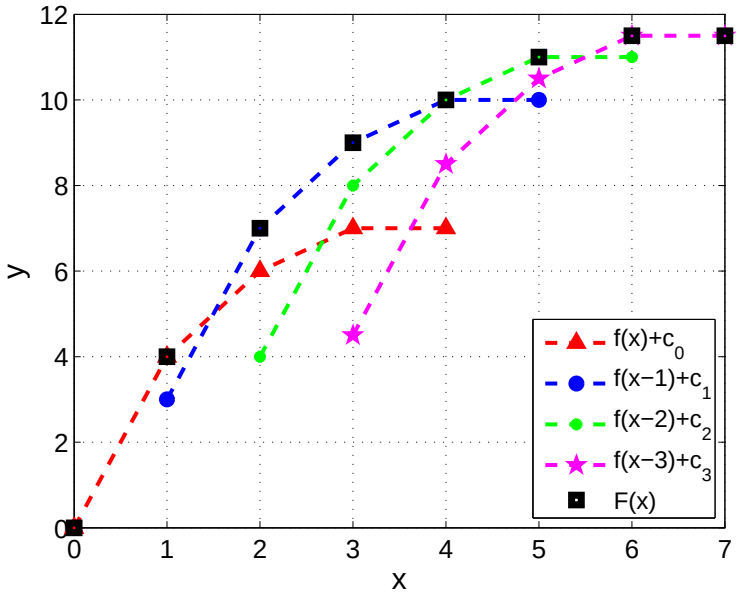


Figure A.1: Illustration of Lemma 2 with $s = 3$.

From (A.19) and (A.20), respectively, we have

$$f(x-(k^*+1))+c_{k^*+1} \geq f(x-(k^*+2))+c_{k^*+2}, \tag{A.21}$$

$$f(x-(k^*-2))+c_{k^*-2} \leq f(x-(k^*-1))+c_{k^*-1}. \tag{A.22}$$

Remark: Based on the same arguments introduced above, it is not difficult to prove that $f(x-k)+c_k$ is monotonically increasing in k when $k \leq k^*$ and monotonically decreasing in k when $k \geq k^*$.

Next, based on the definition of k^* and the above remark, we will prove that

$$F(x-1) = \max\left\{f(x-k^*)+c_{k^*-1}, f(x-k^*-1)+c_{k^*}\right\}, \tag{A.23}$$

$$F(x+1) = \max\left\{f(x-k^*+1)+c_{k^*}, f(x-k^*)+c_{k^*+1}\right\}. \tag{A.24}$$

Notice that, here we assume $f(x-k-1)$ and c_{k+1} have definitions on $k = k^*$. If either of them does not exist, without loss of generality, we can simply discard the corresponding item in the maximum equation in (A.23) and (A.24).

Based on (A.19), we have

$$f(x-(k^*+1))-f(x-(k^*+2)) \geq c_{k^*+1}-c_{k^*}, \tag{A.25}$$

i.e.,

$$f(x-(k^*+1))-c_{k^*} \geq f(x-(k^*+2))+c_{k^*+1}. \tag{A.26}$$

Based on (A.20), we have

$$f(x-(k^*-1))-f(x-k^*) \leq c_{k^*-1}-c_{k^*-2}, \tag{A.27}$$

i.e.,

$$f(x-(k^*-1))+c_{k^*-2} \leq f(x-k^*)+c_{k^*-1}. \tag{A.28}$$

Due to (A.26), (A.28) and the previous remark, we can shown that $f(x-k-1)+c_k$ is monotonically increasing with k when $k \leq k^*-1$ and monotonically decreasing with k when $k \geq k^*$. (A.23) is thus proved.

Similarly, based on (A.19) and (A.20), we can show that

$$\begin{aligned} f(x-k^*)+c_{k^*+1} &\geq f(x-(k^*+1))+c_{k^*+2}, \\ f(x-(k^*-2))+c_{k^*-1} &\leq f(x-(k^*-1))+c_{k^*}, \end{aligned} \tag{A.29}$$

which shows that $f(x-k+1)+c_k$ is monotonically increasing with k when $k \leq k^*$ and monotonically decreasing with k when $k \geq k^*+1$ based on the above remark. (A.24) is thus proved.

In conclusion, now we have

$$\begin{aligned} F(x-1) &= \max\left\{f(x-k^*)+c_{k^*-1}, f(x-k^*-1)+c_{k^*}\right\}, \\ F(x) &= f(x-k^*)+c_{k^*}, \\ F(x+1) &= \max\left\{f(x-k^*+1)+c_{k^*}, f(x-k^*)+c_{k^*+1}\right\}. \end{aligned}$$

If $F(x-1) = f(x-k^*) + c_{k^*-1}$ and $F(x+1) = f(x-k^*) + c_{k^*+1}$, based on the monotonicity and the concavity of c_k, the objective equation (A.16) is true.

If $F(x-1) = f(x-k^*-1) + c_{k^*}$ and $F(x+1) = f(x-k^*+1) + c_{k^*}$, based on the monotonicity and the concavity of $f(x)$, the objective equation (A.16) is true.

If $F(x-1) = f(x-k^*) + c_{k^*-1}$ and $F(x+1) = f(x-k^*+1) + c_{k^*}$, we have

$$\begin{aligned} F(x) - F(x-1) &= c_{k^*} - c_{k^*-1} \geq 0, \\ F(x+1) - F(x) &= f(x-k^*+1) - f(x-k^*) \geq 0. \end{aligned} \tag{A.30}$$

Then, based on (A.18), the objective equation (A.16) is proved.

If $F(x-1) = f(x-k^*-1) + c_{k^*}$ and $F(x+1) = f(x-k^*) + c_{k^*+1}$, we have

$$\begin{aligned} F(x) - F(x-1) &= f(x-k^*) - f(x-k^*-1) \geq 0, \\ F(x+1) - F(x) &= c_{k^*+1} - c_{k^*} \geq 0. \end{aligned} \tag{A.31}$$

Then, based on (A.17), we arrive at (A.16), which completes the proof of Lemma 2.

We now prove [P2] based on Lemma 2. Based on the symmetry property of $N(n,d)$, it is sufficient to consider $d = 0, 1, \ldots, \left\lfloor \frac{n}{2} \right\rfloor$. The proof is based on induction in n using the recursive formulas in (3, 4). The initial condition of the induction is easy to check: $N(2,d)$ is a concave function of d for $0 \leq d \leq 1$. Now assume that there exists an $n_0 > 2$ such that for every $n < n_0$, $\{N(n,d)\}_{d=0}^{\lfloor n/2 \rfloor}$ is a concave sequence in d.

Based on this induction assumption, we prove next that $\{N(n_0,d)\}_{d=0}^{\lfloor n_0/2 \rfloor}$ is a concave sequence in d.

In the following proof, for given n and m, $\phi(m;\, n, d)$ in (2.16) is viewed as a function of d. The maximizer of (2.16) defined as

$$d_1^*(m;\, n, d) := \arg\max_{d_1} \{N(m, d_1) + N(n-m, d-d_1)\}$$

is also viewed as a function of d.

We show next that for given n_0 and m, $\phi(m;\, n_0, d)$ is concave in d. Since $m \le \left\lfloor \frac{n_0}{2} \right\rfloor$, $d \le \left\lfloor \frac{n_0}{2} \right\rfloor$, based on the symmetric property $N(n, d) = N(n, n-d)$ and the increasing property [P1], $d_1^*(m;\, n_0, d)$ must satisfy

$$d - \left\lceil \frac{n_0 - m}{2} \right\rceil \le d_1^*(m;\, n_0, d) \le \left\lfloor \frac{m}{2} \right\rfloor. \tag{A.32}$$

Also since $0 \le d_1^*(m;\, n_0, d) \le d$, we can tighten the range of $d_1^*(m;\, n_0, d)$ to

$$\underline{d_1} \le d_1^*(m;\, n_0, d) \le \overline{d_1},$$

where $\underline{d_1} = \max\left\{0, d - \left\lceil \frac{n_0 - m}{2} \right\rceil\right\}$, $\overline{d_1} = \min\left\{d, \left\lfloor \frac{m}{2} \right\rfloor\right\}$.

Thus $\phi(m;\, n_0, d)$ can be written as

$$\phi(m;\, n_0, d) = \max_{d_1 = \underline{d_1}, \ldots \overline{d_1}} \{N(m, d_1) + N(n_0 - m, d - d_1)\}. \tag{A.33}$$

Note that $N(n_0 - m, d - \underline{d_1})$ is an increasing (based on [P1]) and concave (based on the induction hypothesis) function of d. For the same reasons, $\{N(m, d_1)\}_{d_1 = \underline{d_1}}^{\overline{d_1}}$ is an increasing and concave sequence in d_1. Then Lemma 2 immediately shows that $\phi(m;\, n_0, d)$ is increasing and concave in d, i.e.,

$$2\phi(m;\, n_0, d) \ge \phi(m;\, n_0, d-1) + \phi(m;\, n_0, d+1).$$

We thus have

$$\begin{aligned} &\min_m \{2\phi(m;\, n_0, d)\} \\ &\quad \ge \min_m \{\phi(m;\, n_0, d-1) + \phi(m;\, n_0, d+1)\} \\ &\quad \ge \min_m \{\phi(m;\, n_0, d-1)\} + \min_m \{\phi(m;\, n_0, d+1)\}. \end{aligned} \tag{A.34}$$

Adding 2 to both sides of the inequality, we complete the induction and arrive at [P2].

A.2.3 Proof of [P3]

[P3] can be easily deduced from [P2] as follows.

The condition in [P3] is equivalent to

$$N(m,1) - N(m,0) \leq N(n-m,d) - N(n-m,d-1).$$

Applying the concavity property in [P2] to both sides of this equality leads to

$$N(m,2) - N(m,1) \leq N(n-m,d-1) - N(n-m,d-2),$$

which is equivalent to the statement in [P3] for $d_1 = 2$. Following the same line of argument, we arrive at [P3] for $d_1 = 1, 2, \ldots, \min\{m, d\}$.

APPENDIX B

PROOF FOR THEOREMS IN CHAPTER 3

B.1 Two Sequential Versions of the Local Tests

B.1.1 Passive Sequential Local Test

We now introduce a sequential local test plan based on the Sequential Probability Ratio Testing (SPRT) [81]. For each child of a node on level l, we have two hypotheses which are

$$H_a : \text{the node does not contain the target;}$$

$$H_b : \text{the node contains the target.}$$

The observation distribution of each child node will follow f_{l-1} and g_{l-1} under hypothesis H_a and H_b respectively. Let p_{fp} denote the probability of declaring H_b when H_a is true; and let p_{fn} denote the probability of declaring H_a when H_b is true. The local test first takes samples from the left child. After taking each sample, the SLLR S_L of the left node is updated based on (3.6).

The test of the left node stops as soon as the $S_L > \gamma_1$ or $S_L < \gamma_0$, where

$$\gamma_1 = \log \frac{1 - p_{\text{fn}}}{p_{\text{fp}}}, \quad \gamma_0 = \log \frac{p_{\text{fn}}}{1 - p_{\text{fp}}}. \tag{B.1}$$

The settings of the two thresholds in (B.1) can guarantee any desired p_{fp} and p_{fn} [81]. If $S_L > \gamma_1$, H_1 of the local test is declared, and the IRW policy zooms into the left child. If $S_L < \gamma_0$, the local test declare that the left child does not contain the target (H_a) and switches to the right child. The test on the right child is the same as left child and to detect whether it contains the target with the same setting of thresholds. If the SLLR of

the right child becomes greater than γ_1, the local test declare the right child contains the target. If the SLLR of the right child becomes smaller than γ_0, the local test declare that neither of the two children contains the target. The values of p_{fp} and p_{fp} are chosen to ensure $p_l^{(f)} = (1 - p_{\text{fp}})^2 > \frac{1}{2}$ and $p_l^{(g)} = (1 - p_{\text{fp}})(1 - p_{\text{fn}}) > \frac{1}{2}$.

B.1.2 Active Sequential Local Test

We now present the active version of the local test, which sequentially determines the child to draw the next sample to best distinguish the three hypotheses.

The Sum Log-likelihood ratios (SLLRs), S_L and S_R, of the left and right children nodes are updated separately with their own samples based on (3.6).

The SLLRs of two nodes start from 0. At each time, the active local testing takes one sample from the node which has a larger log-likelihood ratio. If $S_L = S_R$, the sample can be taken from either of the nodes. The active local testing stops as soon as $\max\{S_L, S_R\} \leq \nu_0$, and we then declare H_0 is true; or when $\max\{S_L, S_R\} \geq \nu_1$, and we then declare H_1 is true if $S_L \geq \nu_1$ or H_2 is true if $S_R \geq \nu_1$. When there is only one target in the tree, for the local test, there are three hypotheses.

$$H_0 : \text{neither of the two children contains target;}$$

$$H_1 : \text{the left child contains the target;}$$

$$H_2 : \text{the right child contains the target.}$$

Let p_{ab} denote the probability of declaring hypothesis H_b when H_a is the true one. We now show that the following setting of the thresholds

$$\nu_1 = \log \frac{p_{11}}{p_{01}}, \quad \nu_0 = \log \frac{p_{10}}{p_{00}}, \tag{B.2}$$

can guarantees any desire p_{11} and p_{00} which are required to be greater than $\frac{1}{2}$ for the IRW policy.

Now we show the derivation of the thresholds in (B.2). Let $X_1, X_2, \ldots, ...$ denote the samples taken from the left node and $Y_1, Y_2, \ldots, ...$ denote the samples taken from the right node. Let p_1 and p_0 denote distributions of the samples taken from the node contains or not contains the target, respectively. Define

$$\Lambda_x^k := \prod_{i=1}^{k} \frac{p_1(X_i)}{p_0(X_i)},\ k = 1, 2, \ldots \tag{B.3}$$

$$\Lambda_y^l := \prod_{j=1}^{l} \frac{p_1(Y_i)}{p_0(Y_i)},\ l = 1, 2, \ldots \tag{B.4}$$

which are the likelihood ratios of the left and right child nodes respectively. The *log-likelihood ratios* of there two nodes are defined as

$$S_x^k := \log \Lambda_x^k = \sum_{i=1}^{k} \log \frac{p_1(X_i)}{p_0(X_i)},\ k = 1, 2, \ldots \tag{B.5}$$

$$S_y^l := \log \Lambda_y^l = \sum_{i=1}^{l} \log \frac{p_1(Y_i)}{p_0(Y_i)},\ l = 1, 2, \ldots \tag{B.6}$$

To simplify the notation, let $x := (x_1, \ldots, x_k)$, $y := (y_1, \ldots, y_l)$, and write $p_j(x) = \prod_{i=1}^{k} p_j(x_i)$ and $p_j(y) = \prod_{i=1}^{l} p_j(y_i)$, $j = 0, 1$.

The decision sets of the active local testing can be written as

$$R_0 := \left\{x, y : \Lambda_x^k \le \nu_0,\ \Lambda_y^l \le \nu_0\right\}, \tag{B.7}$$

$$R_1 := \left\{x, y : \Lambda_x^k \ge \nu_1,\ \Lambda_y^l \le 1\right\}, \tag{B.8}$$

$$R_2 := \left\{x, y : \Lambda_x^k \le 1,\ \Lambda_y^l \ge \nu_1\right\}. \tag{B.9}$$

P_{11} can be written in terms of the decision set R_1 as follows:

$$\begin{aligned} P_{11} &= \int_{R_1} p_1(x)p_0(y)dxdy \\ &= \int_{R_1} \frac{p_1(x)}{p_0(x)} p_0(x)p_0(y)dxdy \\ &= \int_{R_1} \Lambda_x^k p_0(x)p_0(y)dxdy \\ &\geq \nu_1 \int_{R_1} p_0(x)p_0(y)dxdy \\ &= \nu_1 P_{01}. \end{aligned} \tag{B.10}$$

P_{00} can be written in terms of the decision set R_0 as

$$\begin{aligned} P_{00} &= \int_{R_0} p_0(x)p_0(y)dxdy \\ &= \int_{R_0} \frac{p_0(x)}{p_1(x)} p_1(x)p_0(y)dxdy \\ &= \int_{R_0} \frac{1}{\Lambda_x^k} p_1(x)p_0(y)dxdy \\ &\geq \frac{1}{\nu_0} \int_{R_0} p_1(x)p_0(y)dxdy \\ &= \frac{1}{\nu_0} P_{10}. \end{aligned} \tag{B.11}$$

Similarly, we can get $P_{22} \geq \nu_1 P_{02}$ and $P_{00} \geq \frac{1}{\nu_0} P_{20}$. In the detection on each level, it is common to set that $P_{11} = P_{22}$, $P_{10} = P_{20}$ and $P_{01} = P_{02}$. These expressions give us bounds on the thresholds necessary to achieve P_{11}, P_{10}, P_{00}, and P_{01}:

$$\nu_1 \leq \frac{P_{11}}{P_{01}} \tag{B.12}$$

$$\nu_0 \geq \frac{P_{10}}{P_{00}} \tag{B.13}$$

We then set

$$\nu_1 = \frac{P_{11}}{P_{01}}, \quad \nu_0 = \frac{P_{10}}{P_{00}}. \tag{B.14}$$

It is easy to see that $P_{11} + P_{10} + P_{12} = 1$, and $P_{00} + P_{01} + P_{02} = P_{00} + 2P_{01} = 1$. In the RWT policy, we require $P_{11} > \frac{1}{2}$ and $P_{00} > \frac{1}{2}$. It can be seen that $\nu_1 > 1$ and $\nu_0 < 1$.

When using the RWT policy, we update the log-likelihood ratios after taking each sample. The test keeps sampling if

$$\log \frac{P_{10}}{P_{00}} < \max\left\{S_x^k, S_y^l\right\} < \log \frac{P_{11}}{P_{01}}.$$

With the proposed active local test, we can also guarantee that $P_{12} \leq P_{02}$, since it is not difficult to see that with the decision set R_2, P_{12} can be written as

$$\begin{aligned} P_{12} &= \int_{R_2} p_1(x)p_0(y)dxdy = \int_{R_2} \frac{p_1(x)}{p_0(x)} p_0(x)p_0(y)dxdy \\ &= \int_{R_2} \Lambda_x^k p_0(x)p_0(y)dxdy \leq \int_{R_2} p_0(x)p_0(y)dxdy = P_{02}. \end{aligned} \tag{B.15}$$

After setting $P_{01} = P_{02}$, we have $P_{12} \leq P_{01}$.

B.1.3 Sequential Local Tests for Multiple Targets Detection

For both the passive and active sequential tests, the testing plan remains the same as the one target case except the updating of the SLLR of the children now is based on

$$\log \frac{h_{l-1}^{(\widehat{d}+1)}(y(n))}{h_{l-1}^{(\widehat{d})}(y(n))}, \tag{B.16}$$

where $\widehat{d}$ is the number of already declared targets that are descendants of the tested child.

Specifically, for the passive sequential test, the left and right children are tested separated. The local test first samples the left child to declare whether is contains any undeclared target(s). If not, it then samples the right child. The stopping thresholds of the SLLR keep the same as in (B.1).

For the multiple targets detection, the active test still samples the child with higher SLLR when $\nu_0 < \max\{S_L, S_R\} < \nu_1$. When $\max\{S_L, S_R\} \leq \nu_0$, the local test declare H_0.

When $\max\{S_L, S_R\} \geq \nu_1$, the test declare the node with larger SLLR have undeclared targets. ν_0 and ν_1 are the same as defined in (B.2).

It can be shown that as long as the family of the distribution $h_l^{(d)}$ satisfies the *Monotone Likelihood-ratio (MLR) Criterion* [54], the active local test is a Uniformly Most Powerful (UMP) test. i.e., the setting of the thresholds for the one target case in (B.1) and (B.2) for the passive sequential and active test, respectively, still work for the multiple targets detection which guarantee any desire probability greater than $\frac{1}{2}$ for the IRW policy, no matter how many undeclared targets exist on the subtree. If the UMP test does not exist for certain distributions, some other local tests for the composite hypothesis testing such as the UMP Invariant Test and Generalized Likelihood Ratio Tests can be considered. As long as the probability of approaching the targets is greater than $\frac{1}{2}$, the IRW policy can guarantee to find the targets with sufficient high probability.

B.2 Proof of Theorem 3

We now give the proof of Theorem 3. Without loss of generality (due to the symmetry of the binary tree structure), we assume that the left-most leaf is the target. We focus on the IRW with fixed-sample local tests.

The random walk on the tree can be divided into two states. The first state is the random walk on upper level nodes of the binary tree. In this state, at each time, after taking K_l samples, we either zoom-in to one child node or zoom-out to the parent node. i.e., the distance between the current node to the target is defined as the sum of the discrete distance to the target node on the tree and the threshold $\log \frac{\log_2 M}{c}$, which will either minus one (zoom-in) or plus one (zoom-out) after every $2K_l$ samples from the children. Once arriving at a leaf node, the test arrives at the second state, where samples

are taken one by one from the current node until the cumulative SLLR exceeds the threshold or becomes negative. The cumulative SLLR can be viewed as a discrete time random walk with random continuous step size which is the log-likelihood ratio of each sample. For all the non-target leaf-nodes, we define the distance between the node to the target as the sum of the discrete distance on the tree, the cumulative SLLR of current node, and the threshold. For the target node, we define the distance to the target as the difference between the threshold and the current cumulative SLLR of the target node. During the search process, these two different states happen consecutively in Phase I of IRW policy.

Let W_n denote the random variable of the step size of the random walk at time n. When the IRW is in the first state (random walk on the high-level nodes), depending on the current level $l > 0$, W_n will have the distribution

$$\Pr(W_n) = \begin{cases} p_l^{(g)} & \text{for } W_n = -1 \\ 1 - p_l^{(g)} & \text{for } W_n = 1 \end{cases} \tag{B.17}$$

if the node is located at a sub-tree contains the target, or

$$\Pr(W_n) = \begin{cases} p_l^{(f)} & \text{for } W_n = -1 \\ 1 - p_l^{(f)} & \text{for } W_n = 1 \end{cases} \tag{B.18}$$

if the node is located at a sub-tree does not contain the target. Since $p_l^{(g)} > 0.5$ and $p_l^{(f)} > 0.5$ for all $l = 1, 2, \ldots, \log_2 M$, we have

$$\mathbb{E}[W_n] = 1 - 2p_l^{(g)} \text{ or } 1 - 2p_l^{(f)},$$

which are both less than 0.

For the second state, let Y_0 and Z_0 denote the random variables with the distributions g_0 and f_0, respectively. The LLR will be either $-\log \frac{g_0(Y_0)}{f_0(Y_0)}$ or $\log \frac{g_0(Z_0)}{f_0(Z_0)}$. It is not difficult

to see that for the target node, we have

$$\mathbb{E}[W_n] = \mathbb{E}\left[-\log \frac{g_0(Y_0)}{f_0(Y_0)}\right] = -D(g_0||f_0) < 0, \tag{B.19}$$

and for all the non-target node, we have

$$\mathbb{E}[W_n] = \mathbb{E}\left[\log \frac{g_0(Z_0)}{f_0(Z_0)}\right] = -D(f_0||g_0) < 0. \tag{B.20}$$

We further assume that the distribution of $-\log \frac{g_0(Y_0)}{f_0(Y_0)}$ and $\log \frac{g_0(Z_0)}{f_0(Z_0)}$ are light-tailed distributions.

Now we are ready to present the following lemma that characterizes the distributions of τ_i.

Lemma 3. *For all τ_i with $i = 1, \ldots, \log_2 M$, there exist an $\alpha > 0$ and a $\gamma > 0$ which are independent of M and c, such that*

$$\Pr(\tau_i \geq n) \leq \alpha e^{-\gamma n}, \ \forall n \geq 0. \tag{B.21}$$

Proof. We first prove this lemma for $\tau_{\log_2 M}$ which is the last passage time of the sub-tree at the root that does not contain the target.

Let S_t denote the distance to the target at time t. The IRW policy starts at the root node, therefore the initial distance to the target is $S_0 = \log_2 M + \log \frac{\log_2 M}{c}$. Define

$$\tau^* = \sup\{t \geq 0 : S_t \geq S_0\} \tag{B.22}$$

as the last time when the search approach has the distance to the target greater than S_0. It is not difficult to see that

$$\tau_{\log_2 M} \leq \tau^*. \tag{B.23}$$

Therefore, we have

$$\Pr(\tau_{\log_2 M} \geq n) \leq \Pr(\tau^* \geq n). \tag{B.24}$$

Based on the definition of τ^*, we have

$$\Pr(\tau^* > n) = \Pr\left(\sup\{t \geq 0 : S_t \geq S_0\} > n\right) \leq \sum_{t=n}^{\infty} \Pr(S_t \geq S_0) = \sum_{t=n}^{\infty} \Pr\left(\sum_{j=1}^{t} W_j \geq 0\right). \tag{B.25}$$

Let μ_j denote the mean value for each W_j, where $\mu_j < 0$ for all $j = 1, 2, \ldots, t$. Applying the Chernoff bound to the sum of independent random variables $\sum_{j=1}^{t} W_j$, for all $s > 0$ we have

$$\Pr\left(\sum_{j=1}^{t} W_j \geq 0\right) \leq \mathbb{E}\left[e^{s\sum_{j=1}^{t} W_j}\right] = \prod_{j=1}^{t} \mathbb{E}\left[e^{sW_j}\right]. \tag{B.26}$$

Note that the moment generating function (MGF) of each W_j is equal to one at $s = 0$. Furthermore, since $\mathbb{E}\left[W_j\right] < 0$ is strictly negative for all $j \geq 1$, differentiating the MGFs of all W_j with respect to s yields strictly negative derivatives at $s = 0$. Because all W_j's are light-tailed distributions, as a result, for all possible distributions of W_j, there exist $s > 0$ and $\gamma > 0$ such that $\mathbb{E}\left[e^{sW_j}\right]$ is strictly less than $e^{-\gamma} < 1$. Hence, from (B.26), we have

$$\Pr\left(\sum_{j=1}^{t} W_j \geq 0\right) \leq e^{-\gamma t}. \tag{B.27}$$

Due to (B.25), we have

$$\Pr(\tau^* > n) \leq \sum_{t=n}^{\infty} \Pr\left(\sum_{i=1}^{t} W_i \geq 0\right) \leq \sum_{t=n}^{\infty} e^{-\gamma t} = \frac{e^{-\gamma n}}{1 - e^{-\gamma}}. \tag{B.28}$$

Let $\alpha = \frac{1}{1-e^{-\gamma}}$, with (B.24), we eventually get Lemma 3 proved for $\tau_{\log_2 M}$.

Because of the recursive definitions of $\tau_1, \tau_2, \ldots, \tau_{\log_2 M}$, the proofs of all the other τ_i will follow the same procedure. □

Based on Lemma 3, we can easily get the following lemma that characterizes the expected value of τ_i.

Lemma 4. *For all τ_i with $i = 1, \ldots, \log_2 M$, there exists a constant $\beta > 0$, such that*

$$\mathbb{E}[\tau_i] \leq \beta. \tag{B.29}$$

Proof. Based on the the tail-sum formula of expectation of the non-negative random variables, we have

$$\mathbb{E}[\tau_i] = \sum_{n=0}^{\infty} \Pr[\tau_i > n] \leq \sum_{n=0}^{\infty} \alpha e^{-\gamma n} = \frac{\alpha}{1 - e^{-\gamma}} = \frac{1}{(1 - e^{-\gamma})^2} = \beta. \tag{B.30}$$

□

Now we are ready to prove Theorem 3. Base on Lemma 4, it is not difficult to show that

$$\mathbb{E}[\tau] \leq 2K_{\max} \sum_{i=1}^{\log_2 M} \mathbb{E}[\tau_i] + \mathbb{E}[\tau_0] \leq 2\beta K_{\max} \log_2 M + \mathbb{E}[\tau_0]. \tag{B.31}$$

When the observations are informative at all levels, $K_{\max}$ is bounded by a constant, the first term in (B.31) is upper bounded by $B \log_2 M$, where B is a constant greater than $2\beta K_{\max}$.

For the last stage, τ_0 is a stopping time with respect to the i.i.d. sequence of the log-likelihood ratio $\left\{\log \frac{g_0(X_n)}{f_0(X_n)} : n \geq 1\right\}$, where X_n denote i.i.d. random variable with distribution g_0.

Due to the Wald's Equation [81], we have

$$\mathbb{E}\left[\sum_{n=1}^{\tau_0} \log \frac{g_0(X_n)}{f_0(X_n)}\right] = \mathbb{E}[\tau_0]\mathbb{E}\left[\log \frac{g_0(X_n)}{f_0(X_n)}\right]. \tag{B.32}$$

i.e.,

$$\log \frac{\log_2 M}{c} + R_b = \mathbb{E}[\tau_0] D(g_0 \| f_0), \tag{B.33}$$

where R_b is the overshooting at the threshold. Due to Lorden's inequality [64], we have

$$\mathbb{E}[R_b] \leq \frac{\mathbb{E}\left[\left(\log \frac{g_0(X_n)}{f_0(X_n)}\right)^2\right]}{\mathbb{E}\left[\log \frac{g_0(X_n)}{f_0(X_n)}\right]}. \tag{B.34}$$

Assuming that the first two moments of log-likelihood ration are finite, then we have

$$\mathbb{E}[\tau_0] = \frac{\log \frac{\log_2 M}{c}}{D(g_0 \| f_0)} + O(1). \tag{B.35}$$

The following lemma characterizes the error probability of the IRW policy.

Lemma 5. *The error probability of the IRW policy is upper bounded by:*

$$P_e \leq \beta c = O(c). \tag{B.36}$$

Proof. When the IRW policy arrives a non-target node, say node j, the probability of error (accepting H_j) equals to $\Pr(S_j \geq \log \frac{\log_2 M}{c})$. The Wald's approximation [82] gives

$$\Pr(S_j \geq \log \frac{\log_2 M}{c}) \leq \exp\left[-\log \frac{\log_2 M}{c}\right] = \frac{c}{\log_2 M}. \tag{B.37}$$

Let N denote the random number of times of visiting these non-target leaf nodes in the IRW policy. The conditional error probability is upper bounded by $\frac{Nc}{\log_2 M}$. Based on the proof of Theorem 3, the expected value of N is upper bounded by $\beta \log_2 M$. Therefore, by taking expectation, the error probability is bounded by

$$P_e \leq \frac{c}{\log_2 M} \cdot \mathbb{E}[N] \leq \frac{c}{\log_2 M} \cdot \beta \log_2 M = \beta c = O(c). \tag{B.38}$$

□

Based on (B.31), (B.35) and Lemma 5, we thus arrive at Theorem 3.

B.3 Proof of Theorem 4 and Theorem 5

In this appendix, we provide the proof of Theorem 4 and Theorem 5. Follow the same lines of the argument, we still have the sample complexity of the last stage τ_0 satisfy (B.35). We now give the upper bound of the sample complexity of the first $\log_2 M$ stages for the test with fixed-size local test.

We focus on the Bernoulli distribution model, where g_l and f_l are Bernoulli distribution with false negative and false positive rates equal to μ_l. In order to get the relation between K_l and μ_{l-1}, we first introduce the following lemma [66].

Lemma 6. *Let $X_1, \ldots, X_n$ be independent Poisson trails such that $\Pr(X_i) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\nu = \mathbf{E}[X]$. Then the following Chernoff bounds hold for $0 < \delta \leq 1$,*

$$\Pr(X \geq (1+\delta)\nu) \leq e^{-\nu\delta^2/3}; \tag{B.39}$$

$$\Pr(X \leq (1-\delta)\nu) \leq e^{-\nu\delta^2/3}. \tag{B.40}$$

By applying Lemma 6, it is not difficult to show in order to have $p_l^{(g)}$ and $p_l^{(f)}$ define in (3.7) greater than 0.5 for all $l = 1, 2, \ldots, \log_2 M$, we can choose K_l greater than

$$\max\left\{\frac{12(1-\mu_{l-1})\log(1-\eta)^{-1}}{(1-2\mu_{l-1})^2}, \frac{12\mu_{l-1}\log(1-\lambda)^{-1}}{(1-2\mu_{l-1})^2}\right\}, \tag{B.41}$$

where η and λ can be any value in $(\frac{1}{\sqrt{2}}, 1)$ such that $\eta \cdot \lambda > 0.5$ and $\lambda^2 > 0.5$. Since $\mu_l < 0.5$, w.l.o.g., we choose

$$K_l = \frac{12(1-\mu_{l-1})\log(1-\eta)^{-1}}{(1-2\mu_{l-1})^2}. \tag{B.42}$$

It is not difficult to find that K_l increases with μ_{l-1}. For any stage l, when $l = 1, 2, \ldots, \log_2 M$, the sample complexity in this stage is upper bounded by $2K_l \cdot \mathbb{E}[\tau_l]$. Due to Lemma 4, the total sample complexity from Stage 1 to Stage $\log_2 M$ is thus upper bounded by

$$\mathbb{E}[\tau] \leq \sum_{l=1}^{\log_2 M} 2K_l \cdot \mathbb{E}[\tau_l] \leq \sum_{l=1}^{\log_2 M} 2\beta K_l. \tag{B.43}$$

For Theorem 4, if $\mu_l = 0.5 - (0.5 - \mu_0) \cdot (l+1)^{-\alpha}$, due to (B.42) and (B.43), we have

$$\mathbb{E}[\tau] \leq B' \sum_{l=1}^{\log_2 M} l^{2\alpha}, \tag{B.44}$$

where $B' = \frac{6\beta\log(1-\eta)^{-1}}{(0.5-\mu_0)^2}$ is a constant. By using the Faulhaber's formula, we have

$$\sum_{l=1}^{\log_2 M} l^{2\alpha} = O((\log_2 M)^{2\alpha+1}).$$

Thus Theorem 4 is proved.

Similarly, for Theorem 5, if $\mu_l = 0.5 - (0.5 - \mu_0) \cdot \alpha^{-l}$, we have

$$\mathbb{E}[\tau] \leq B' \sum_{l=1}^{\log_2 M} \alpha^{2(l-1)}. \tag{B.45}$$

By summing up the geometric terms in (B.45), we can show that

$$\mathbb{E}[\tau] \leq \tilde{B}(\alpha^2)^{\log_2 M} = \tilde{B} M^{\frac{2}{\log_2 \alpha^2}}, \tag{B.46}$$

where $\tilde{B} = \frac{1}{\alpha^2 - 1} B'$. Thus Theorem 5 is proved.

B.4 Proof of Theorem 6

To prove Theorem 6, we first show the sample complexity of the IRW policy satisfies

$$\mathbb{E}\left[\tau | \Gamma_{\text{IRW}}\right] \leq LB \log_2 M + \frac{L \log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + o(L \log \frac{M}{c}). \tag{B.47}$$

In the one-target detection proof, we defined the random walk as the distance from the current testing node to the target. Because of the random walk is biased, i.e., the IRW policy guarantees the probability of approaching the target is always greater than 0.5, the expectation of each step of the random walk is always approaching the target. By using Chernoff bound, we show that the last passage time $\mathbb{E}[\tau_l]$ on the tree $\mathscr{T}_l$ for all $l = 1, 2, \ldots, \log_2 M$ is upper bounded by a constant. Then sample complexity in the first $\log_2 M$ stages are in logarithmic-order. For the last stage on $\mathscr{T}_0$, it can be shown that $\mathbb{E}[\tau_0] = \frac{\log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + O(1)$.

The basic idea to prove (B.47) is similar to the one target case. For the multiple targets detection, we need to find a proper random variable that defines a random walk. The desired property of the random variable is that the expectation of the changing of

the random variable is negative at each step of the random walk. i.e., as the IRW policy approaching the targets, the random variable has the trend to keeps decrease to zero until the targets been found. Then by using the Chernoff bound, we would be able get the similar upper bound to the detection delay for the multiple-targets case.

The IRW policy is designed to find the targets one by one. As the process going, there may be detection errors in the previous round. We split the proof of the overall detection delays into two cases, which are the detection delays with or without detection errors on the tree respectively.

B.4.1 Detection delay without detection errors on the tree

Different with the one targets detection, we modify the definition of $\mathscr{T}_l$ for all $l = 1, 2, \ldots, \log_2 M$ for the multiple targets detection case. As illustrated in Fig. 3.5 for $M = 8$ and $L = 3$, our approach is to partition the tree into $\log_2 M + 1$ disjoint sets of nodes. Similar to the one target case, the detection process of finding any one of the targets is then partitioned into $\log_2 M + 1$ stages by the successively defined last passage time to each of the set of nodes from upper level to lower level.

We now give the proof of the detection delay of finding the first target. The random walk on the tree also has two states. The first state is the random walk on the upper level nodes of the binary tree. Once arriving at a leaf node, the test arrives at the second state, where samples are taken one by one from the current node until the cumulative SLLR exceeds the threshold or becomes negative. Without loss of generality, we numerate all the targets with index 1 to L from left to right. For any node v on the tree, we define $D_{\min}(v)$ as

$$D_{\min}(v) := \min_{i=1,\ldots,L} \{D_i(v)\}, \tag{B.48}$$

where $D_i(v)$ is the distance on the tree between current node v to the ith target.

For all the non-target leaf-nodes v, we define the distance between the node to the target as the sum of $D_{\min}(v)$, the cumulative SLLR of current node, and the threshold $\log \frac{\log_2 M}{c}$. For the target node, we define the distance to the target as the difference between the threshold and the current cumulative SLLR of the target node.

Let W_n denote one step of the global random walk at time n. When the IRW is in the first state, given the current node v, $W_n = \Delta D_{\min}(v)$ can be either 1 or -1, which has the distribution

$$\Pr(W_n) = \Pr(\Delta D_{\min}(v)) = \begin{cases} p_l(v) & \text{for } W_n = \Delta D_{\min}(v) = -1; \\ 1 - p_l(v) & \text{for } W_n = \Delta D_{\min}(v) = 1. \end{cases} \tag{B.49}$$

In IRW policy, for a node v on level l, after taking K_l samples, the random walk has probability $p_l(v)$ greater than $\frac{1}{2}$ to approach the targets in the tree rooted at current node or have $p_l(v) > \frac{1}{2}$ to zoom out of the current node if it contains no targets. Therefore, we have

$$\mathbb{E}[W_n] = \mathbb{E}[\Delta D_{\min}(v)] = 1 - 2p_l(v) < 0,$$

which shows that, by applying IRW policy, the expectation of the each step of the random is to approach at least one of the targets on the tree.

In the second state, similar to the one target detection process, we have (B.19) for the target nodes and (B.20) for all the non-target leaves.

Similarly, let τ_i denote the lass passage time to set $\mathscr{T}_i$. More specifically, τ_i is also the last time that the random walk have distance greater or equal to $i + \log \frac{\log_2 M}{c}$ to all the targets. i.e., after τ_i, the random walk will have distance less than $i + \log \frac{\log_2 M}{c}$ to at least one of the targets. Then use the same arguments in the proof of one target detection, we

have for all τ_i with $i = 1, \ldots, \log_2 M$, there exists a constant $\beta > 0$, such that

$$\mathbb{E}[\tau_i] \leq \beta. \tag{B.50}$$

Therefore, the detection delay $\mathbb{E}[\tau]$ of finding a target in the first round is upper bounded by

$$E[\tau] \leq B \log_2 M + \frac{\log \frac{\log_2 M}{c}}{D(g_0 \| f_0)} + O(1). \tag{B.51}$$

Similarly, the error probability is bounded by

$$P_e \leq \beta c = O(c).$$

For the subsequent $L - 1$ rounds to finding the remaining $L - 1$ targets, as long as there are no detection errors happen in the detection, the detection delay of each round can be bounded as in (B.51). Applying the union bound, the overall probability of error is bounded by

$$P_e \leq L\beta c = O(Lc). \tag{B.52}$$

Therefore, with probability at least $1 - O(Lc)$, the detection delay of finding all the L targets is upper bounded by

$$E[\tau_{\text{all}}] \leq LB \log_2 M + \frac{L \log \frac{\log_2 M}{c}}{D(g_0 \| f_0)} + O(L). \tag{B.53}$$

B.4.2 Detection delay with detection errors on the tree

We now show the detection delay when there are detection errors on the tree. Assume there are total L targets remaining to be detected and there are total E detection errors. Due to the detection errors on the tree, the preference of the IRW policy to approaching the targets may changes on part of the tree.

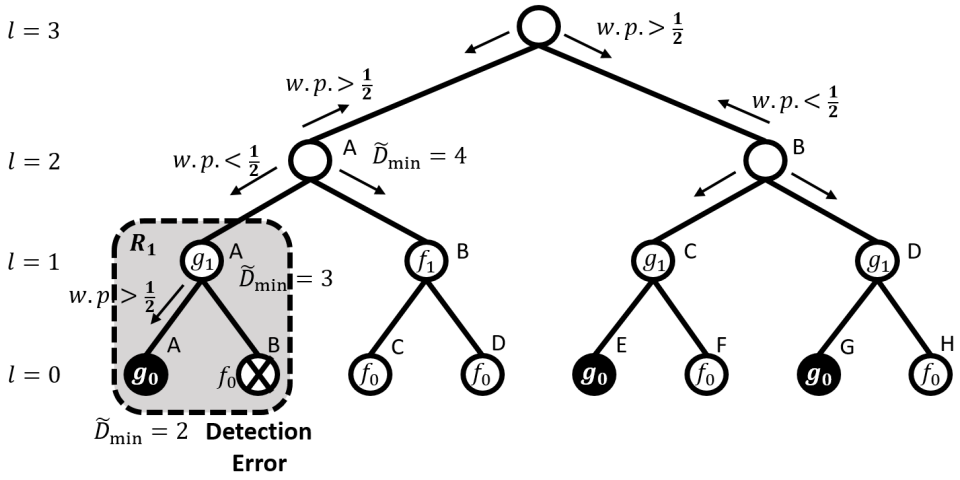


Figure B.1: A biased random walk on the tree with detection errors.

In Fig. B.1, we illustrate an example with $M = 8$, $L = 3$, and $E = 1$. Assume that after the first round of the test, there is an detection error happened on level $l = 0$ node B. In the next round of test, when applying the IRW policy and starting from the root node, the probability of approaching the two targets on the right half tree is always greater than 0.5. However, on the left half tree, due to the detection error, the observation on the higher level nodes would make the decision maker think that there is no more undeclared targets on the left half tree. The probability of approaching the left-most target is less than $\frac{1}{2}$ before the random walk entering the subtree R_1 as shown in Fig. B.1. But once the random walk enters the subtree R_1, the probability of approaching the left-most target becomes greater than $\frac{1}{2}$ since the detection error will not affect the observation from the true target anymore. In the other word, for all the nodes below node A on level $l = 1$, the random walk will have higher probability of approaching the left most target; for all the nodes above node A on level $l - 1$, the random walk will have higher probability of leaving the left target. We call R_1 as the *affected subtree*, the node A on level $l = 1$ as *changing point*, and the left-most target as the *affected target*.

For the general case, we provide the definition of these terminologies as follows. Since the affected trees may be in a nested structure, they are defined in a recursive way.

Definition 2. *If an undeclared target has a detection error as sibling, the subtree formed by these two nodes and their parent node will be an affected subtrees. Starting from the lower level to upper level, any minimum subtrees that contain at least one undeclared target node which is not covered by any other lower level affected trees and contains more or equal number of declaration errors than the undeclared targets are also called affected subtrees.*

Definition 3. *Roots of the affected subtrees are called changing points.*

Definition 4. *All the undeclared targets in an affected subtree are called affected targets.*

There may be multiple affected subtrees in the detection and they are possibly in a nested structure. We illustrate another example in Fig. B.2, where R_1 and R_2 are two affected trees in a nested structure.

Our objective is to show the detection delay of the IRW policy is upper bound when there are detection errors in the tree. The proof idea is similar as before, we need to find a proper random variable that has negative expectation (approaching the targets) at each step of the random walk.

Let $\mathcal{V}$ denote the set of all the target nodes; C denote the set of all targets that have already been correctly declared; $\mathcal{A}$ denote the set of the undeclared targets which are affected by the declaration erros; $\mathcal{U}$ denote the set of the undeclared targets which are not affected by the declaration errors. It is easy to see that C, $\mathcal{A}$ and $\mathcal{U}$ are disjoint and $\mathcal{V} = C \cup \mathcal{A} \cup \mathcal{U}$.

For any node v on the tree, depending on whether the node is on an affected tree, we consider the following two cases.

If v is not on any affected trees Define

$$\tilde{D}_{\min}(v) := \min_{i \in \mathcal{U}} \{D_i(v)\}, \tag{B.54}$$

which is the minimum distance on the tree from v to the undeclared targets which are not affected by the declaration errors.

If v is on an affected tree Since the affected trees may be in a nested structure, $\tilde{D}_{\min}(v)$ can be defined in a recursive way. Let v_c denote the changing point of the affected subtree and D_c denote the minimum distance from the change point to the undeclared targets on this affected tree.

We define $\tilde{D}_{\min}(v)$ for the node v from larger affected subtrees to smaller affected subtrees, from higher level to lower level. For the highest level changing point v of the largest affected subtree, the parent of v must not be on any affected trees, of which the $\tilde{D}_{\min}$ is defined in the previous bullet. We define a constant Z as

$$Z := \tilde{D}_{\min}(\text{parent node of } v_c) - D_c - 1. \tag{B.55}$$

It is not difficult to see that $Z \geq 0$.

Within all the nodes on current affected subtree which are not covered by any lower level nested subtrees, let $\mathcal{V}_R$ and $\mathcal{V}_T$ denote the sets of all tree nodes and all the undeclared targets, respectively. For any node $v \in \mathcal{V}_R$, $\tilde{D}_{\min}(v)$ is defined as

$$\tilde{D}_{\min}(v) = Z + \min_{i \in \mathcal{V}_T}\{D_i\}. \tag{B.56}$$

For the nodes on all the lower level/nested affected trees, we use the (B.55) and (B.56) recursively to find $\tilde{D}_{\min}(v)$. It is not difficult to see that if there are no detection errors on the tree, $\tilde{D}_{\min}(v)$ will coincide with $D_{\min}(v)$ defined in (B.48).

We now apply the definitions in (B.55) and (B.56) to some examples. As shown in Fig B.1, $\tilde{D}_{\min}(v)$ of node A on level $l = 2$ is 4 based on (B.54). For the affected subtree R_1, Z equals 2. Therefore, $\tilde{D}_{\min}(v)$ of node A on level $l = 1$ is 3. For the example in Fig B.2, there are two affected subtrees R_1 and R_2. For R_1, Z equals 6. Therefore, $\tilde{D}_{\min}(v)$ for the node A on level $l = 3$ is 9 and for the node A on $l = 2$ is 10. For R_2, Z equals 8, which makes $\tilde{D}_{\min}(v)$ for the node A on level $l = 1$ be 9.

It is not difficult to see that after each step of the random walk variable $W_n = \Delta D_{\min}(v)$ will have the distribution

$$\Pr(W_n) = \Pr\left(\Delta\tilde{D}_{\min}(v)\right) = \begin{cases} p_l(v) & \text{for } W_n = \Delta\tilde{D}_{\min}(v) = -1; \\ 1 - p_l(v) & \text{for } W_n = \Delta\tilde{D}_{\min}(v) = 1. \end{cases} \tag{B.57}$$

The IRW policy guarantees that $p_l(v)$ is always greater than 0.5. Therefore, we have

$$\mathbb{E}[\Delta\tilde{D}_{\min}(v)] = 1 - 2p_l(v) < 0.$$

Similar to the sample complexity without detection errors, let τ_i denote the last time that the random walk have distance greater or equal to $i + \log \frac{\log_2 M}{c}$ to all the targets. i.e., after τ_i, the random walk will have distance less than $i + \log \frac{\log_2 M}{c}$ to at least one of the targets. However, due to the definition the maximum value of $\tilde{D}_{\min}$ can be as large as $2\log_2 M$. Use the same arguments in the proof of one target detection, we have for all τ_i with $i = 1, \ldots, 2\log_2 M$, there exists a constant $\beta > 0$, such that

$$\mathbb{E}[\tau_i] \leq \beta. \tag{B.58}$$

Because of the add-on constant Z in the definition of $\tilde{D}_{\min}$ in (B.56), the first state of the random walk may stop before $\sum_{i=1}^{2\log_2 M} t_i$. In this case, the detection delay of the random walk on the first state will still be bounded. Therefore, when there are detection

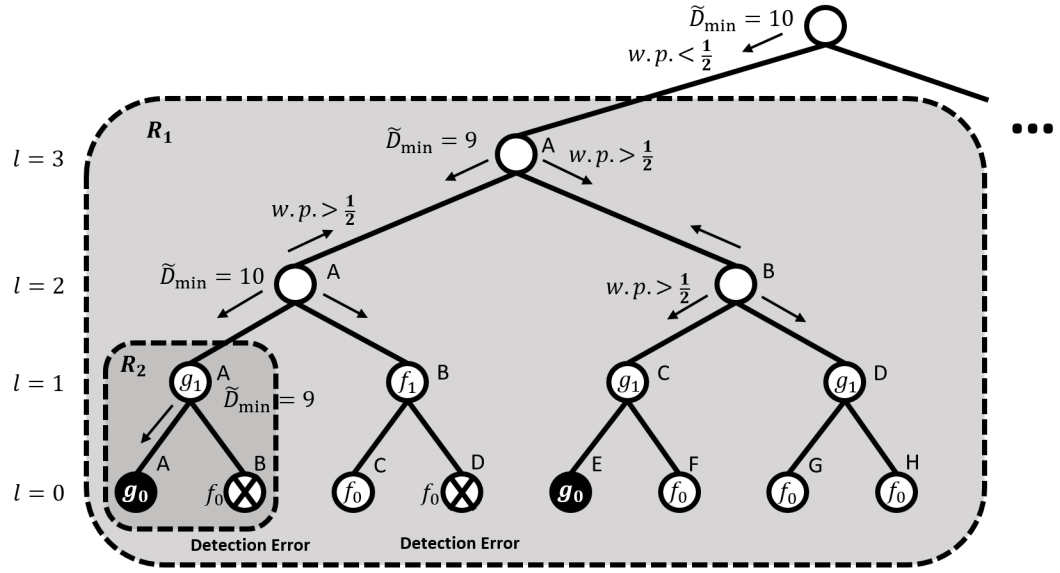


Figure B.2: A biased random walk on the tree with detection errors: nested affected trees.

errors on the tree, the detection delay $\mathbb{E}[\tau]$ of finding a target is upper bounded by

$$E[\tau] \leq 2B \log_2 M + \frac{\log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + O(1).$$

With probability at most $O(Lc)$, the detection delay of finding all the L targets when there are detection errors happened is upper bounded by

$$E[\tilde{\tau}_{\text{all}}] \leq 2LB \log_2 M + \frac{L \log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + O(L). \tag{B.59}$$

Combining (B.53) and (B.59), we have the detection delay of applying the IRW policy to the L-targets detection problem be upper bounded by

$$\begin{aligned} \mathbb{E}\left[\tau | \Gamma_{\text{IRW}}\right] &\leq (1 - L\beta c) E[\tau_{\text{all}}] + L\beta c E[\tilde{\tau}_{\text{all}}] \\ &\leq LB \log_2 M + \frac{L \log \frac{\log_2 M}{c}}{D(g_0 || f_0)} + L^2 B \beta c \log_2 M. \end{aligned} \tag{B.60}$$

Based on (B.52) and (B.60), Theorem 6 can be therefore proved.

BIBLIOGRAPHY

[1] M Aigner and M Schughart. Determining defectives in a linear order. *Journal of Statistical Planning and Inference*, 12:359–368, 1985.

[2] Martin Aigner. Search problems on graphs. *Discrete Applied Mathematics*, 14(3):215–230, 1986.

[3] Matthew Aldridge, Leonardo Baldassini, and Oliver Johnson. Group testing algorithms: Bounds and simulations. *IEEE Transactions on Information Theory*, 60(6):3671–3687, 2014.

[4] George K Atia and Venkatesh Saligrama. Boolean compressed sensing and noisy group testing. *IEEE Transactions on Information Theory*, 58(3):1880–1901, 2012.

[5] DJ Balding, WJ Bruno, DC Torney, and E Knill. A comparative survey of non-adaptive pooling designs. In *Genetic mapping and DNA sequencing*, pages 133–154. Springer, 1996.

[6] T. Berger, N. Mehravari, D. Towsley, and J. Wolf. Random multiple-access communication and group testing. *IEEE Transactions on Communications*, 32(7):769–779, Jul 1984.

[7] Toby Berger and Vladimir I Levenshtein. Asymptotic efficiency of two-stage disjunctive testing. *IEEE Transactions on Information Theory*, 48(7):1741–1749, 2002.

[8] Radu Berinde, Anna C Gilbert, Piotr Indyk, Howard Karloff, and Martin J Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 798–805, 2008.

[9] Stuart Alan Bessler. Theory and applications of the sequential design of experiments, k-actions and infinitely many experiments: Part I–Theory. *Tech. Rep. Applied Mathematics and Statistics Laboratories, Stanford University*, (55), 1960.

[10] Nader H Bshouty. Optimal algorithms for the coin weighing problem with a spring scale. In *COLT*, 2009.

[11] Nader H Bshouty and Hanna Mazzawi. Toward a deterministic polynomial time algorithm with optimal additive query complexity. In *Mathematical Foundations of Computer Science*, pages 221–232. Springer, 2010.

[12] Marat Valievich Burnashev. Data transmission over a discrete channel with feedback. random transmission time. *Problemy peredachi informatsii*, 12(4):10–30, 1976.

[13] Sheng Cai, Mohammad Jahangoshahi, Mayank Bakshi, and Sidharth Jaggi. Grotesque: noisy group testing (quick and efficient). In *51st Annual Allerton Conference on Communication, Control, and Computing*, pages 1234–1241. IEEE, 2013.

[14] Rui Castro and Robert Nowak. Active learning and sampling. In *Foundations and Applications of Sensor Management*, pages 177–200. Springer, 2008.

[15] Chun Lam Chan, Sidharth Jaggi, Venkatesh Saligrama, and Samar Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. *IEEE Transactions on Information Theory*, 60(5):3019–3035, 2014.

[16] Shih-Chun Chang and E Weldon. Coding for T-user multiple-access channels. *IEEE Transactions on Information Theory*, 25(6):684–691, 1979.

[17] M. Cheraghchi, A. Hormati, A. Karbasi, and M. Vetterli. Group testing with probabilistic tests: Theory, design and application. *IEEE Transactions on Information Theory*, 57(10):7057–7067, Oct 2011.

[18] M. Cheraghchi, A. Karbasi, S. Mohajer, and V. Saligrama. Graph-constrained group testing. *IEEE Transactions on Information Theory*, 58(1):248–262, Jan 2012.

[19] Mahdi Cheraghchi. Derandomization and group testing. In *48th Annual Allerton Conference on Communication, Control, and Computing*, pages 991–997, 2010.

[20] Mahdi Cheraghchi, Ali Hormati, Amin Karbasi, and Martin Vetterli. Compressed sensing with probabilistic measurements: A group testing solution. In *47th Annual Allerton Conference on Communication, Control, and Computing*, pages 30–35. IEEE, 2009.

[21] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959.

[22] Sung-En Chiu and Tara Javidi. Sequential measurement-dependent noisy search. In *Information Theory Workshop (ITW)*, pages 221–225. IEEE, 2016.

[23] Sung-Soon Choi and Jeong Han Kim. Optimal query complexity bounds for finding graphs. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 749–758. ACM, 2008.

[24] Cisco. Sampled NetFlow. `http://www.cisco.com/c/en/us/td/docs/ios/12_0s/feature/guide/12s_sanf.html`.

[25] Kobi Cohen and Qing Zhao. Active hypothesis testing for anomaly detection. *IEEE Transactions on Information Theory*, 61(3):1432–1450, 2015.

[26] Giacomo Como, Fabio Fagnani, and Sandro Zampieri. Anytime reliable transmission of real-valued information through digital noisy channels. *SIAM Journal on Control and Optimization*, 48(6):3903–3924, 2010.

[27] Peter Damaschke. Threshold group testing. In *General theory of information transfer and combinatorics*, pages 707–718. Springer, 2006.

[28] AG Djackov. On a search model of false coins. In *Topics in Information Theory (Colloquia Mathematica Societatis Janos Bolyai 16, Keszthely, Hungary). Budapest, Hungary: Hungarian Acad. Sci*, page 163170, 1975.

[29] Robert Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[30] Ding-Zhu Du and Frank K Hwang. *Combinatorial group testing and its applications*, volume 12. World Scientific, 1999.

[31] Dingzhu Du and Frank Hwang. *Combinatorial group testing and its applications*. World Scientific, 2nd edition, 2000.

[32] Dingzhu Du and Frank Hwang. *Pooling Design and Nonadaptive Group Testing: Important Tools for DNA Sequencing*. World Scientific, 2006.

[33] Amin Emad and Olgica Milenkovic. Semiquantitative group testing. *IEEE Transactions on Information Theory*, 60(8):4614–4636, 2014.

[34] Paul Erdös, Peter Frankl, and Zoltán Füredi. Families of finite sets in which no set is covered by the union of others. *Israel Journal of Mathematics*, 51(1):79–89, 1985.

[35] Paul Erdos and Alfréd Rényi. On two problems of information theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl*, 8:229–243, 1963.

[36] Cristian Estan and George Varghese. *New directions in traffic measurement and accounting*, volume 32. ACM, 2002.

[37] W. Fang and L. Peterson. Inter-AS traffic patterns and their implications. In *Global Telecommunications Conference*, volume 3, pages 1859–1868, 1999.

[38] NI Fine. Solution of problem E 1399. *American Mathematical Monthly*, 67(7):697–698, 1960.

[39] Luisa Gargano, V Montouri, G Setaro, and Ugo Vaccaro. An improved algorithm for quantitative group testing. *Discrete applied mathematics*, 36(3):299–306, 1992.

[40] Jun Geng, Weiyu Xu, and Lifeng Lai. Quickest search over multiple sequences with mixed observations. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pages 2582–2586, 2013.

[41] Allen Gersho and Robert M Gray. *Vector quantization and signal compression*, volume 159. Springer Science & Business Media, 2012.

[42] Anna C Gilbert, Mark A Iwen, and Martin J Strauss. Group testing and sparse signal recovery. In *42nd Asilomar Conference on Signals, Systems and Computers*, pages 1059–1063, 2008.

[43] Vladimir Grebinski and Gregory Kucherov. Optimal reconstruction of graphs under the additive model. In *Algorithms-ESA'97*, pages 246–258. Springer, 1997.

[44] Vladimir Grebinski and Gregory Kucherov. Optimal reconstruction of graphs under the additive model. *Algorithmica*, 28(1):104–124, 2000.

[45] Weidong Han, Peter I Frazier, and Bruno M Jedynak. Probabilistic group testing under sum observations: A parallelizable 2-approximation for entropy loss. *arXiv preprint arXiv:1407.4446*, 2015.

[46] Fred H Hao. The optimal procedures for quantitative group testing. *Discrete Applied Mathematics*, 26(1):79–86, 1990.

[47] Jarvis Haupt, Richard Baraniuk, Rui Castro, and Robert Nowak. Sequentially designed compressed sensing. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 401–404, 2012.

[48] Jarvis Haupt, Rui M Castro, and Robert Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 57(9):6222–6235, 2011.

[49] Michael Horstein. Sequential transmission using noiseless feedback. *IEEE Transactions on Information Theory*, 9(3):136–143, 1963.

[50] Piotr Indyk and Milan Ruzic. Near-optimal sparse recovery in the l1 norm. In *49th Annual IEEE Symposium on Foundations of Computer Science*, pages 199–207, 2008.

[51] Mark A Iwen and Ahmed H Tewfik. Adaptive strategies for target detection and localization in noisy environments. *IEEE Transactions on Signal Processing*, 60(5):2344–2353, 2012.

[52] Sina Jafarpour, Weiyu Xu, Babak Hassibi, and Robert Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Transactions on Information Theory*, 55(9):4299–4308, 2009.

[53] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, 2008.

[54] Samuel Karlin and Herman Rubin. The theory of decision procedures for distributions with monotone likelihood ratio. *The Annals of Mathematical Statistics*, pages 272–299, 1956.

[55] Gyula OH Katona. Combinatorial search problems. *A survey of combinatorial theory*, pages 285–308, 1973.

[56] W Kautz and Roy Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

[57] W Kautz and Roy Singleton. Nonrandom binary superimposed codes. *IEEE Transactions on Information Theory*, 10(4):363–377, 1964.

[58] S. Khattab, S. Gobriel, R. Melhem, and D. Mosse. Live baiting for service-level DoS attackers. In *The 27th IEEE Conference on Computer Communications*, April 2008.

[59] Tze Leung Lai. Nearly optimal sequential tests of composite hypotheses. *The Annals of Statistics*, pages 856–886, 1988.

[60] Anusha Lalitha, Nancy Ronquillo, and Tara Javidi. Improved target acquisition rates with feedback codes. *arXiv preprint arXiv:1712.05865*, 2017.

[61] Anusha Lalitha, Nancy Ronquillo, and Tara Javidi. Measurement dependent noisy search: The gaussian case. In *IEEE International Symposium on Information Theory (ISIT)*, pages 3090–3094, 2017.

[62] Chou Hsiung Li. A sequential method for screening experimental variables. *Journal of the American Statistical Association*, 57(298):455–477, 1962.

[63] B Lindström et al. Determining subsets by unramified experiments. 1975.

[64] Gary Lorden. On excess over the boundary. *The Annals of Mathematical Statistics*, pages 520–527, 1970.

[65] Matthew L Malloy and Robert D Nowak. Near-optimal adaptive compressed sensing. *IEEE Transactions on Information Theory*, 60(7):4001–4012, 2014.

[66] Michael Mitzenmacher and Eli Upfal. *Probability and computing: Randomized algorithms and probabilistic analysis*. Cambridge University Press, 2005.

[67] Mohammad Naghshvar and Tara Javidi. Active sequential hypothesis testing. *The Annals of Statistics*, 41(6):2703–2738, 2013.

[68] Hung Q Ngo and Ding-Zhu Du. A survey on combinatorial group testing algorithms with applications to DNA library screening. *Discrete mathematical problems with medical applications*, 55:171–182, 2000.

[69] Sirin Nitinawarat, George K Atia, and Venugopal V Veeravalli. Controlled sensing for multihypothesis testing. *IEEE Transactions on Automatic Control*, 58(10):2451–2464, 2013.

[70] Jonathan Scarlett and Volkan Cevher. Converse bounds for noisy group testing with arbitrary measurement matrices. In *International Symposium on Information Theory (ISIT)*, number EPFL-CONF-215128, 2016.

[71] Harold S Shapiro. Problem E 1399. *Amer. Math. Monthly*, 67(82):697–697, 1960.

[72] A. Sharma and C. Murthy. Group testing based spectrum hole search for cognitive radios. *IEEE Transactions on Vehicular Technology*, PP(99):1–1, 2014.

[73] Albert N Shiryaev. *Optimal stopping rules*, volume 8. Springer Science & Business Media, 2007.

[74] Milton Sobel and Phyllis A Groll. Group testing to eliminate efficiently all defectives in a binomial sample. *Bell System Technical Journal*, 38(5):1179–1252, 1959.

[75] Ali Tajer, Venugopal V Veeravalli, and H Vincent Poor. Outlying sequence detection in large data sets: A data-driven approach. *IEEE Signal Processing Magazine*, 31(5):44–56, 2014.

[76] Vincent YF Tan and George Atia. Strong impossibility results for noisy group testing. In *ICASSP*, pages 8257–8261, 2014.

[77] My Tra Thai, Ying Xuan, Incheol Shin, and Taieb Znati. On detection of malicious users using group testing techniques. In *The 28th International Conference on Distributed Computing Systems*, pages 206–213, 2008.

[78] K. Thompson, G.J. Miller, and R. Wilder. Wide-area internet traffic patterns and characteristics. *IEEE Network*, 11(6):10–23, Nov 1997.

[79] Sattar Vakili, Qing Zhao, Chang Liu, and Chen-Nee Chuah. Anomaly detection in hierarchical data streams under unknown models. *arXiv preprint arXiv:1709.03573*, 2017.

[80] Rolf Waeber, Peter I Frazier, and Shane G Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, 2013.

[81] A. Wald. Sequential tests of statistical hypotheses. *Ann. Math. Statist.*, 16(2):117–186, 06 1945.

[82] Abraham Wald. *Sequential analysis.* John Wiley, 1947.

[83] C. Wang, Q. Zhao, and C. N. Chuah. Optimal nested test plan for combinatorial quantitative group testing. *IEEE Transactions on Signal Processing*, 66(4):992–1006, Feb 2018.

[84] J. Wolf. Born again group testing: Multiaccess communications. *IEEE Transactions on Information Theory*, 31(2):185–191, Mar 1985.

[85] Jack Keil Wolf. Principles of group testing and an application to the design and

analysis of multi-access protocols. In *The Impact of Processing Techniques on Communications*, pages 237–257. Springer, 1985.

[86] Weiyu Xu and Babak Hassibi. Efficient compressive sensing with deterministic guarantees using expander graphs. In *IEEE Information Theory Workshop*, pages 414–419, 2007.

[87] Weiyu Xu and Babak Hassibi. Further results on performance analysis for compressive sensing using expander graphs. In *41st Asilomar Conference on Signals, Systems and Computers*, pages 621–625, 2007.

[88] Minlan Yu, Lavanya Jose, and Rui Miao. Software defined traffic measurement with OpenSketch. In *NSDI*, volume 13, pages 29–42, 2013.